Partially-Synchronized DEC-MDPs in Dynamic Mechanism Design

Partially-Synchronized DEC-MDPs in Dynamic Mechanism Design

Sven Seuken, Ruggiero Cavallo, and David C. Parkes
School of Engineering and Applied Sciences
Harvard University
Cambridge, MA 02138
{seuken,cavallo,parkes}@eecs.harvard.edu

Abstract

In this paper, we combine for the first time the methods of dynamic mechanism design with techniques from decentralized decision making under uncertainty. Consider a multi-agent system with self-interested agents acting in an uncertain environment, each with private actions, states and rewards. There is also a social planner with its own actions, rewards, and states, acting as a coordinator and able to influence the agents via actions (e.g., resource allocations). Agents can only communicate with the center, but may become inaccessible, e.g., when their communication device fails. When accessible to the center, agents can report their local state (and models) and receive recommendations from the center about local policies to follow for the present period and also, should they become inaccessible, until becoming accessible again. Without self-interest, this poses a new problem class which we call *partially-synchronized DEC-MDPs*, and for which we establish some positive complexity results under reasonable assumptions. Allowing for *self-interested* agents, we are able to bridge to methods of dynamic mechanism design, aligning incentives so that agents truthfully report local state when accessible and choose to follow the prescribed "emergency policies" of the center.

Introduction

Imagine a scenario in which there is a taxi company with taxis that have private state information (e.g., their location in the city) which they can report to central dispatch who cannot observe their state. The taxis have *private actions* (e.g., driving a specific route) and receive private rewards (e.g., payment by passengers net fuel cost), both of which are again unobservable by the center. Central dispatch has information about customers waiting at designated locations and can assign them to particular taxis (i.e., making a resource allocation decision). Furthermore, the center can suggest specific routes or locations to the taxis.

Taxi drivers are *self-interested* and seek to maximize individual utility, which depends on local state, local actions, and also the center's resource allocation decisions. This can lead to incentive problems, for instance with many taxi drivers claiming they are closest to a passenger wishing to go to the airport. We explore in this paper the use of dynamic mechanisms, wherein the center can provide payments (both to and from taxis) in order to align the local goals of each driver with those of the entire company.

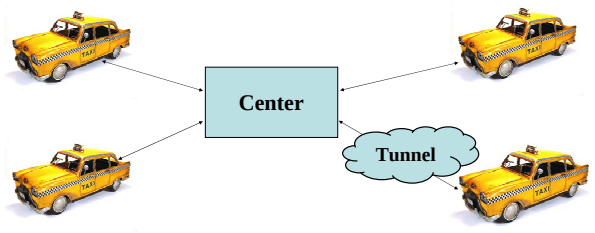

Figure 1: Taxi domain: coordination of self-interested agents with private actions and periodic inaccessibility.

A particularly interesting challenge arises when taxis might become *inaccessible* for certain time periods, for example because they are driving through a tunnel or because their communication device fails. During times of inaccessibility the center continues taking central actions and inaccessible agents continue taking local actions, but they cannot communicate and payments cannot be made. See Figure 1.

The combination of both unobservable and thus not "contractible" actions and periodic inaccessibility makes this a novel problem in *dynamic mechanism design*, where one seeks to align incentives with the social good (Parkes 2007). We model the goal of the center as that of a "social planner", interested in implementing sequential, system-optimal decisions, both at the local and central levels, despite conflicting interests of the individual agents. To address our setting (with private actions), we need to extend the methods of Cavallo et al. (2007) and Bergemann & Välimäki (2006).

Without inaccessibility, the decision problem can be modeled as a *multi-agent MDP* (Boutilier 1999), which can be solved in time polynomial in the input size of the problem. At every time step the center could elicit state reports and then compute a new joint action. With inaccessibility and private actions, on the other hand, this is precluded.

Now, in every time step, the center takes an action and proposes local policies for each accessible agent to follow in the current period and also in the future should the agent become inaccessible. The resulting computational problem is related to the family of *decentralized Markov decision processes* (DEC-MDPs, DEC-POMDPs) and significantly harder to solve than MMDPs. However, the problem we study—that of *partially-synchronized DEC-MDPs*—is quite

different from existing DEC-MDP models in that agents are only periodically inaccessible and local agents can "synchronize" their local states with the center when accessible.

An interesting aspect of this work is the dance between incentive requirements and computational requirements: we are able to carve out a class of models that avoids the complexity of many problems in the DEC-MDP class while also facilitating the alignment of agent incentives.

Related Work

Cavallo et al. (2006) introduced an efficient (i.e., value-maximizing), incentive-compatible mechanism for an environment in which agents have a dynamically evolving local state. Bergemann & Välimäki (2006) independently developed the *dynamic-VCG* mechanism that is able to achieve a stronger participation constraint, so that agents will want to continue to participate in the mechanism whatever the current state. Cavallo et al. (2007) provide an extension that allows for *periodically inaccessible* agents. But none of these earlier mechanisms allow for the coupling of both local actions and periodic inaccessibility. We content ourselves here presenting a dynamic Groves mechanism that aligns incentives but runs at a deficit and defer presenting the full dynamic-VCG generalization to the full paper. Athey and Segal (2007) consider a similar domain (without inaccessibility) and devise an efficient mechanism that is budget-balanced but with weaker participation properties. See Parkes (2007) for a recent survey.

Bernstein et al. (2002) introduced the DEC-MDP model, the first formal framework for decentralized decision making under uncertainty. It and all subsequent models have assumed a context of cooperative agents. At each time step, an agent takes a local action, a state transition occurs, and each agent receives a local observation from the environment. Thus, each agent can have different partial information about the other agents and about the state of the world. Bernstein et al. (2002) proved that the finite-horizon version of this problem is NEXP-complete, and thus is likely to require double exponential time to solve in the worst case. Becker et al. (2004) show that with transition- and observation-independence the finite-horizon version of the problem is NP-complete. But this model is insufficient for our purposes because we have periodic communication opportunities between the center and agents and also because it would preclude an agent from observing a reward that depends on the center's action. See Seuken and Zilberstein (2008) for a survey.

Partially-Synchronized DEC-MDPs

We now introduce our new model. Consider a multi-agent domain with $n+1$ agents: a set of agents $I = \{1, \ldots, n\}$ and a designated center agent. We refer to agents $1, ..., n$ as *local agents* and we use the term "agents" when referring to local agents and the center agent. In our domain, the center plays a special role because he coordinates all local agents.

Definition 1 (PS-DEC-MDP). *A partially-synchronized DEC-MDP is a tuple* $\langle I, \{S_i\}, \{AS_i\}, S_c, s^0, \{A_i\}, A_c, \{P_i\}, P_c, \{R_i\}, R_c, T, \gamma \rangle$.

- Each agent $i \in \{1, \ldots, n\}$ has a set of local states S_i. A characteristic feature of our domain is that local agents can become inaccessible to the center. The set of of *accessible states* for agent i is denoted by $AS_i \subseteq S_i$.
- The center agent has a finite set of center states S_c. The space of local states and the center's state space constitute the global system state space $S = S_c \times S_1 \times S_2 \times \ldots \times S_n$. Let $s^0 = \langle s_c^0, s_1^0, s_2^0, ..., s_n^0 \rangle$ denote the initial system state. We require that all agents are initially accessible, i.e., $s_i^0 \in AS_i$ for all $i \in I$.
- Each local agent has a set of local actions A_i and the center has a set of center actions A_c.
- The transitions and rewards of local agents can depend on the center but not on other local agents. Let $P_i : S_i \times S_c \times A_i \times A_c \rightarrow \Delta S_i$ denote the transition function for agent i, where $P_i(s_i'|s_i, s_c, a_i, a_c)$ denotes the probability that, after taking actions a_i and a_c in states s_i and s_c, a transition to local state s_i' occurs. Let $R_i : S_i \times S_c \times A_i \times A_c \rightarrow \Re$ denote the reward function for agent i.
- For inaccessible states, agent i's transition and reward is also independent of the center, i.e. there exists P_i' s.t. $\forall s_i \in S_i \setminus AS_i, s_c \in S_c, a_c \in A_c : P_i(s_i'|s_i, s_c, a_i, a_c) = P_i'(s_i'|s_i, a_i)$. Similarly, for the reward there exists some R_i' s.t. $R_i(s_i, s_c, a_i, a_c) = R_i'(s_i, a_i)$ for all $s_i \in S_i \setminus AS_i, s_c \in S_c, a_c \in A_c$.
- The center's transition function is $P_c : S_c \times A_c \rightarrow \Delta S_c$ and completely independent of local agents. The center's intrinsic reward $R_c : S_c \times A_c \rightarrow \Re$ for its own actions has the same independence.[1]
- The problem can either have a finite or infinite horizon. A finite horizon is denoted with a positive integer T, and for infinite horizons T is replaced with ∞.
- γ is a discount factor $\in [0, 1]$.

Both the local agents and the center take actions and receive rewards, but we impose the particular independence structure so that local agent problems are independent of those of other agents when we condition on the action and state of the center. The inaccessibility of an agent is an attribute of the local state space of an agent. When all agents are accessible we say that the system is "synchronized."

The PS-DEC-MDP model does not model communication actions explicitly. However, in solving the model we will allow for the agents to communicate their local model and local state to the center when accessible, and for the center to communicate a local policy back to each agent. An agent is unable to communicate with the center when inaccessible and an agent's reward and transitions are also completely independent while inaccessible.[2]

[1] This is not to be confused with the higher-level goal of the center as a social planner, which is then to design a system that maximizes the sum of its own intrinsic reward and that of the agents. In this paper, we will assume the center's intrinsic reward is always 0.

[2] This independence when inaccessible is crucial for complexity considerations because it ensures that when an agent is inaccessible there is no way any information can be transmitted between the

Policies and Value Functions for PS-DEC-MDPs

To complete the PS-DEC-MDP model we need to define both the center's policy and the policy of local agents. For this we first introduce semantics for *observations* in this domain. Each local agent and the center always observes its own local state. But because agents can go inaccessible, the state of the system *as viewed by the center* is not always the true system state $s^t \in S$, but rather that reflected by the most recent reports from agents. The planning problem facing the center must allow for this partial information.

At every time step t, the center observes its own local state $s_c \in S_c$ and receives information $s_i \in AS_i \cup \{\varepsilon\}$ from every agent i, where ε modeling missing information an agent. The center also remembers the last time that agent was accessible (or reported to be accessible). We will generally assume that there is a parameter L_i, the *maximal limit of inaccessibility for agent* i. Thus, at any time step, the inaccessibility information for the center about agent i will be $l_i \in \mathcal{L}_i = \{0, 1, \ldots, L_i - 1\}$, where 0 corresponds to being accessible right now.

But for optimal decision making, the center does not need to remember this whole stream of information/observations he gets over time. To account for this fact we will slightly abuse the term "observation" and define the *center's observation space* as $O_c = AS_1 \times \mathcal{L}_1 \times AS_2 \times \mathcal{L}_2 \times \ldots \times AS_n \times \mathcal{L}_n \times S_c$. We will say that at every time step t, the center is in a new "observation state" $o_c^t \in O_c$. Based on that and the current time t, the center 1) determines action $\pi_c(o_c^t, t) = a_c^t$ for himself, and 2) computes for each agent i new *local policies* $\pi_i : S_i \times T \rightarrow A_i$ of length L_i. Note that the observation state o_c^t is a sufficient statistic for the past stream of observations of the center and can thus be used for optimal meta-level decision making. Due to space constraints, we omit a formal proof for this which would work analogous to the proof of Lemma 2 in (Goldman and Zilberstein 2004).

These policies prescribe to the agents which action to take given local state and time for the next L_i time steps forward, should they become inaccessible. Thus, they can be seen as "*emergency policies*." The decision policy for the center at time t is thus a *meta-decision policy*: $\boldsymbol{\pi}(o_c^t, t) = \langle \pi_c, \pi_1, ..., \pi_n \rangle$.[3] Because all agents are accessible in initial state s^0, every agent gets an initial local policy. For notational purposes we will assume that $\pi_i^t = \pi_i^{t-1}$ if agent i is not accessible at time t. This means that local policy π_i in the meta-decision policy vector stays the same as long as agent i is inaccessible.

It bears emphasis that this joint policy $\boldsymbol{\pi}$ is not implemented on joint state $s \in S$ but rather implemented by each agent (including the center) on its observation state. In fact, it is convenient to adopt o_i^t to denote agent i's local observation state, which is simply $o_i^t = s_i^t$. Again, the local observation state o_i^t is a sufficient statistic for the past history of true observations for agent i and can thus be used for optimal local decision making. With this, then $o^t = \langle o_c^t, o_1^t, ..., o_n^t \rangle = \langle o_c^t, s_1^t, ..., s_n^t \rangle$ denotes the *joint observation state* at time t. Note that o^t contains the information about the true system state and we will write $s(o^t)$ to denote that. Given meta-decision policy $\boldsymbol{\pi}(o_c^t, t)$, we denote the joint action in period t (again, with fully cooperative agents) as $A(o^t, \boldsymbol{\pi}(o_c^t)) = \langle \pi_c(o_c^t), \pi_1(o_1^t), ..., \pi_n(o_n^t) \rangle = \langle a_c, a_1, ..., a_n \rangle = a$.

In any time step t, given current system state s^t and joint action a^t, the joint reward from the social planner's perspective is $R(s^t, a^t) = \sum_{i=1}^{n} R_i(s_i^t, s_c^t, a_i^t, a_c^t) + R_c(s_c^t, a_c^t)$. Solving a PS-DEC-MDP means finding the center's meta-decision policy $\boldsymbol{\pi}^* \in \boldsymbol{\Pi}$ that maximizes the expected total reward over the problem horizon. We can define the value of a meta-decision policy $\boldsymbol{\pi}$ for a finite-horizon PS-DEC-MDP (the infinite-horizon case is analogous) with initial system state s^0 as:

$$V^{\boldsymbol{\pi}}(s^0) = E\Big[\sum_{t=0}^{T-1} \gamma^t R(s^0, a^0) | s^0, \boldsymbol{\pi}\Big].$$

The center's observation state together with meta-decision policy $\boldsymbol{\pi}$ is sufficient to compute a distribution over system states. Thus, we can also define the value of $\boldsymbol{\pi}$ for o_c^t:

$$V^{\boldsymbol{\pi}}(o_c^t) = P(o^t | o_c^t, \boldsymbol{\pi}) \cdot E\Big[\sum_{t=0}^{T-1} \gamma^t R(s(o^t), A(o^t, \boldsymbol{\pi}(o_c^t))) | o^t, \boldsymbol{\pi}\Big].$$

We define the optimal value for observation state o_c^t as:

$$V^*(o_c^t) = \max_{\boldsymbol{\pi} \in \boldsymbol{\Pi}} V^{\boldsymbol{\pi}}(o_c^t)$$

Computational Complexity of PS-DEC-MDPs

Note that for the complexity analysis we always consider the decision problem for PS-DEC-MDPs. That is, given the problem description and a threshold value K, is there a policy for the problem with expected value greater than K? Furthermore, we make the following two assumptions:

Assumption 1. *For finite-horizon problems, we assume that the time horizon T is polynomially bounded by the size of the problem description. This is a standard assumption which is also required to show P-completeness of MDPs.*

Assumption 2. *We assume that the number of agents is a fixed constant. This is necessary to achieve positive complexity results because the observation space of the center is already exponential in the number of agents.*

Theorem 1. *Finite-horizon partially-synchronized DEC-MDPs without a priori limits on inaccessibility are NP-complete.*

Proof. To show NP-hardness we reduce the NP-complete problem DTEAM (Papadimitriou and Tsitsiklis 1986) to a finite-horizon PS-DEC-MDP with a center and one local agent. The proof is similar to one given by (Becker et al. 2004) but our proof is more involved because we require independent reward functions when agent 1 is inaccessible.

center and that agent, implicitly or explicitly. Note also that the local agents as well as the center can always observe their rewards after each time step. This is different from most of the existing models (like the DEC-MDP model), but seems essential to capturing the essence of domains with self-interest.

[3]Note that the center only actually has to determine its own action for the current state, but we adopt the convention that it determines its own (base) policy π_c for convenience.

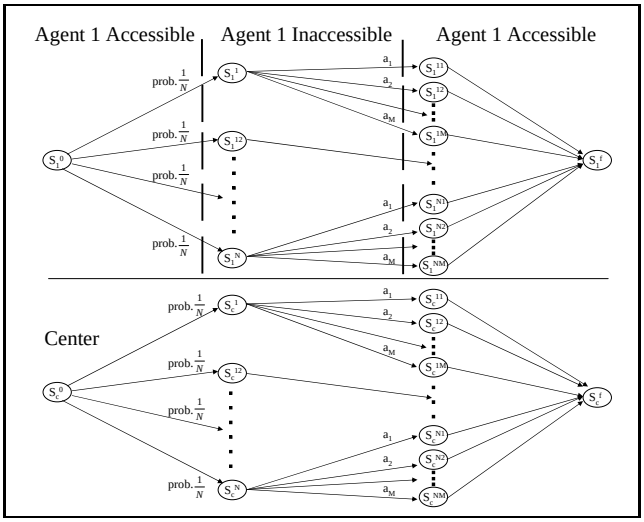

Figure 2: Reduction of the DTEAM problem (NP-complete) to a finite-horizon PS-DEC-MDP.

DTEAM is a decentralized team decision problem. Two agents independently observe random integers $k_1, k_2 \in \{1, ..., N\}$. Then they have to choose actions $\gamma_1(k_1), \gamma_2(k_2) \in \{1, ..., M\}$. A joint cost function $c(k_1, k_2, a_1, a_2)$ specifies the resulting cost for every tuple of observed integers and action pairs taken. The optimal solution to this problem specifies local policies γ_1, γ_2 such that the expected joint cost is minimized.

The reduction is illustrated in Figure 2. The agents start in states s_1^0, s_c^0 and agent 1 is accessible. Then, with uniform probability, agent 1 transitions to a new local state $s_1^i, i \in \{1, ..., N\}$ which is an inaccessible state. The center also transitions to a new local state $s_c^i, i \in \{1, ..., N\}$. Then, both agents can choose a local action $\pi_1(s_1), \pi_c(s_c) \in \{a_1, ..., a_M\}$ based on their local state which results in a new state transition as shown in Figure 2. Now agent 1 is accessible again and the reward function for agent 1 is defined based on the joint state and joint action for the last time step the following way:

$$R(s_1^{ij}, s_c^{lm}, *, *) = -c(k_i, k_l, a_j, a_m)$$

where $*$ indicates that any action can be taken

For the other time steps the reward is always 0. It is now easy to see that any joint policy $\langle \pi_c, \pi_1 \rangle$ that maximizes the expected reward for the resulting PS-DEC-MDP also constitutes a joint policy for the original DTEAM problem minimizing the expected cost. This completes the reduction.

To prove membership in NP we first show that the representation size of the center's meta-decision policy is polynomial in the input size of the problem. As explained in the discussion on the policy representation, the center's observation state o_c^t is a sufficient statistic for optimal decision making. Note that for a finite-horizon PS-DEC-MDP, the time horizon T is an implicit limit on each agent's maximal time of inaccessibility. Thus, in this case, the observation space for the center is $O_c = AS_1 \times T \times AS_2 \times T \times \cdots \times AS_n \times T \times S_c$ which can be represented in size $O(|AS_i|^n \cdot log(T)^n)$. Because the number of agents n is considered fixed, this is polynomial in the problem size. The meta-decision policy is a mapping from observation states and current time to local policies. It is easy to see that the representation size for the local policies is smaller than O_c and thus also polynomially bounded. Thus, the representation size of the center's meta-decision policy is polynomial in the size of the problem. We can guess an optimal meta-decision policy and evaluate it in polynomial time which shows membership in NP. □

Theorem 2. *Finite-horizon partially-synchronized DEC-MDPs with a priori limits $L_1, L_2, ..., L_n$ on the maximal time of inaccessibility are P-complete.*

Proof. P-hardness follows because an MMDP (P-complete) can be reduced to a PS-DEC-MDP without inaccessible agents. Membership in P follows from the limits on the time of inaccessibility. With the limits L_i we know that the maximal number of different observation states o_c the center can be in is $K = |AS_i|^n \cdot L_1 \cdot L_2 \cdot \ldots \cdot L_n \cdot |S_c|$. Again, this number is exponential in the number of agents but otherwise polynomial. We can see the K observation states as "meta-states" for the center, thus, the center's meta-decision policy can now be described as a mapping from meta states and current time T to policy vectors. Now, for all local agents i the local policies are of maximum length L_i. The policy for the center, π_c, can be a one-step policy. Thus, both policy representations are polynomial in the size of the problem description. For every observation state o_c and every time t, we can enumerate all local policies of length L_i (length 1 for the center). Obviously, this is a number of policies exponential in the L_i's but this is not a problem for the complexity analysis because we have assumed that the limits L_i are known a priori and are not part of the problem description.

Once we have enumerated all policies those can be treated as meta-actions and we can use dynamic programming to compute the optimal policy in time polynomial in the size of the problem description. Thus, we can decide in polynomial time whether there exists a policy whose value exceeds threshold K which shows membership in P. □

Theorem 3. *Infinite-horizon PS-DEC-MDPs without limits on inaccessibility are undecidable.*

Proof. To show this result we will leverage a proof by (Madani 2000) who showed that infinite-horizon UMDPs (unobservable MDPs) are undecidable. Madani constructs a 3-state UMDP with 2 actions where both actions a and b executed in any state lead to the same stochastic state transition. However, their reward vectors $R_a()$ and $R_b()$ for the 3 states are different. Madani shows that the optimal policy for that UMDP is an infinite non-periodic action sequence. This result allows him to prove undecidability.

We can reduce this 3-state UMDP to a PS-DEC-MDP with a center and just 1 local agent. The resulting PS-DEC-MDP is depicted in Figure 3. Agent 1 has 8 states where for 3 of them he is inaccessible and for the other 5, he is accessible. The center just has three states where s_a and s_b correspond to the actions a and b in the UMDP. Both agents' local state s_0 is simply the initial start state where agent 1 is still accessible. Agent 1 has no actions but still makes

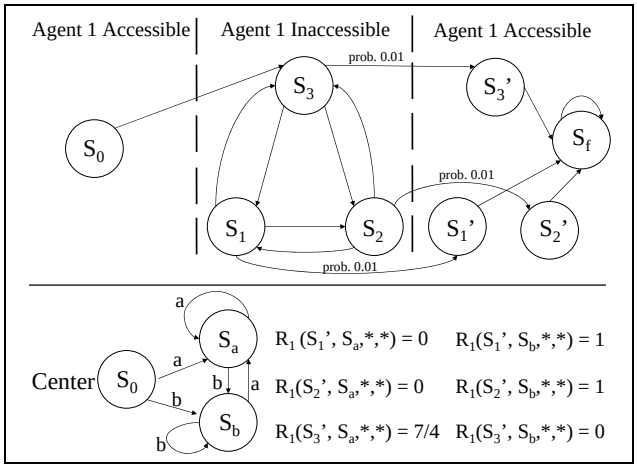

Figure 3: Reduction of an undecidable infinite-horizon UMDP to an infinite-horizon PS-DEC-MDP.

stochastic state transitions. Once the agent is in states s_1, s_2 or s_3, with probability 0.99 agent 1 follows the state transition matrix outlined by Madani. But in each state s_i, with probability 0.01 agent 1 transitions to the corresponding accessible state s_i'. From there, agent 1 transitions to a final absorbing state s_f. Thus, agent 1's transitions are independent of any actions taken. The center can take action a or b in every time step which lead to a deterministic transition to the corresponding state. However, the center does not know agent 1's state as long as agent 1 is inaccessible.

The center never gets a reward and only when agent 1 is in one of the three accessible states s_1', s_2', s_3', he gets a positive reward that depends on his state and the center's state but is independent of any actions (see Figure 3). The rewards are set up such that Madani's reward vector for action a corresponds to the reward vector when the center is in state s_a. Analogous, the reward vector for action b corresponds to the reward vector when the center is in state s_b.

The center's problem is that he must try to be in the reward maximizing state when agent 1 becomes accessible. Thus, the center must optimize its action sequence while agent 1 is inaccessible based on the sequence of belief points in the 3-point simplex (s_1, s_2, s_3). As long as agent 1 is inaccessible, the corresponding sequence of optimal actions (a vs. b) is irregular, as was shown by Madani. It is now easy to see that the optimal local policy for the center π_c for this PS-DEC-MDP would also constitute the optimal policy for Madani's UMDP. Because infinite-horizon UMDPs are undecidable this shows that infinite-horizon PS-DEC-MDPs are undecidable. □

As before for the finite-horizon case, the complexity changes when we put a limit on the time of inaccessibility. In particular, now the problem becomes decidable.

Theorem 4. *Infinite-horizon partially synchronized DEC-MDPs with limits $L_1, L_2, ..., L_n$ on the maximal time of inaccessibility are NP-complete (when the limits are part of the problem description and polynomially bounded by the size of the problem description).*

Proof. We show NP-hardness by reducing the DTEAM problem described in the proof for Theorem 1 to an infinite-horizon PS-DEC-MDP. We use the same reduction as shown in Figure 2 with two small modifications. First, the agents can continue taking actions when they have reached the final state, however, the state stays the same and no rewards are collected. Second, discount factor γ is set to any value in $(0, 1)$. Because the agents only collect rewards at a single time step, γ has no influence on the optimal policy. Thus, the policy that is optimal for the resulting infinite-horizon PS-DEC-MDP is also optimal for the DTEAM problem.

To show membership in NP we we can leverage the well-known fact that there exist stationary optimal policies for infinite-horizon MDPs, where stationary means the optimal policy is a mapping from states to actions and independent of time. With the time limits L_i we know that an upper limit on the number of different observation states o_c the center can observe is $K = |AS_i|^n \cdot L_1 \cdot L_2 \cdot ... \cdot L_n \cdot |S_c|$. Again, this number is exponential in the number of agents but we consider the number of agents to be a fixed constant. We can see those K observation states as "meta-states" for the center, thus, the center's meta-decision policy can now be described as a mapping from meta-states to policy vectors and in particular, this mapping is independent of time. For all local agents i their policies can be of maximum length L_i and the center's policy can be a one-step policy. Thus, both policy representations are polynomial in the problem size. We can guess the optimal meta-decision policy and verify that its value exceeds threshold K. All this can be done in polynomial time, which shows membership in NP. □

Theorem 5. *Infinite-horizon partially-synchronized DEC-MDPs with a priori limits $L_1, L_2, ..., L_n$ on the maximal time of inaccessibility are P-complete.*

Proof. P-completeness follows easily by reducing an MMDP to a PS-DEC-MDP without inaccessible agents. Membership in P can be proved by appealing to the proofs for Theorems 2 and 4. We argue again that we will have stationary policies on a finite number of meta-states. Then, we can enumerate all possible policies, evaluate them (e.g., via linear programming), and verify that the value of the best meta-decision policy exceeds threshold K. This can be done in polynomial time, which shows membership in P. □

Incentive Analysis: Self-Interested Agents

We now introduce self-interest on the part of agents, and address how to leverage the PS-DEC-MDP framework within the context of dynamic incentive mechanisms. The goal is to implement the socially optimal policy subject to the constraints on information provided by the periods of inaccessibility. We present a dynamic mechanism that will provide incentives for self-interested agents to truthfully participate, both in the sense of truthfully reporting local state (and local models) when accessible and also to faithfully follow the prescribed local policies of the center.

In every period, the mechanism elicits reports about state from accessible agents and makes payments to all agents that sent such a report (and are therefore accessible). We adopt the common assumption of *quasilinear utility* so that

an agent's net payoff in a state is the sum of its intrinsic reward and the payment made by the center. Given space constraints we present only a "Groves mechanism" and defer the presentation of the appropriate variant on the dynamic-VCG# mechanism (Bergemann & Välimäki 2006; Cavallo, Parkes, and Singh 2007) for an extended version of the paper. For now this means that the mechanism we present makes the system a "team game" and runs at a deficit.

Modeling Agent Strategies

At each time step t, the center wants all accessible agents to report their local state, and all agents (accessible or not) to follow their prescribed local policies. In order to formalize the equilibrium notion we need to be explicit about an agent's knowledge, independence assumptions, and communication structures. For his local strategy, agent i has to reason about all agents in the system. But due to the communication structure we imposed, agent i can only possibly collect information about the other agents when he is accessible and only then about the other accessible agents. *Thus, from the constraints of the PS-DEC-MDP environment, and in equilibrium when other agents are truthful, we can assume that agent i can never know more about the other agents than the center.* Because the center's observation state o_c contains all information necessary to compute an expectation over future system states s, it is sufficient therefore for agent i to form a belief over the center's observation state; we let $bo_i \in \Delta O_c$ denote a distribution on the center's observation space.[4]

To capture the strategic problem facing an agent we introduce some further notation. Let $\bar{S}_i$ denote the space of vectors of agent i's local states since i was last accessible. Π_i is the space of local agent policies that can be prescribed by the center. Given the following three informational items $bo_i \in \Delta O_c, \bar{s}_i \in \bar{S}_i$ and $\pi_i \in \Pi_i$, agent i can compute his belief over the whole system. Thus, we can define his complete belief space $\mathcal{B}_i = \Delta O_c \times \bar{S}_i \times \Pi_i$ (where we require consistency between bo_i and π_i). Note that a belief state $b_i \in \mathcal{B}_i$ is not simply a distribution over system states, but a more complex informational structure.

For local agents' strategies we first allow for a *reporting strategy* $f_i : \mathcal{B}_i \times T \rightarrow AS_i \cup \{\varepsilon\}$. When agent i is accessible, it can a) report its state truthfully, b) report a false state, or c) pretend to be inaccessible (which we will denote as state report ε). Furthermore, agent i plays an *action strategy* $g_i : \mathcal{B}_i \times T \rightarrow A_i$. In every time period t, for any belief state b_i, $g_i(b_i, t)$ denotes the action taken by the local agent. Let $f = (f_1, \ldots, f_n)$ and $g = (g_1, \ldots, g_n)$ denote the reporting strategy profile and the action strategy profile for the system of n agents. An agent is *truthful* if it always reports true state information to the center when accessible, never pretends to be inaccessible when it is accessible, and abides by the center's policy prescriptions.

[4]To correctly update his belief, agent i would also need the stochastic models of all agents including the center. However this notion of belief is in essence a device we need for the proof, and we will establish that truthful behavior is an equilibrium for *all* beliefs and thus the agent will not need to do this kind of belief-update reasoning; as such, the agent does not require these stochastic models.

Definition 2 (Mechanism). *A mechanism $M = (\boldsymbol{\pi}, X)$ in our environment is defined by a meta-policy $\boldsymbol{\pi}$ for the PS-DEC-MDP along with a payment function $X(o_c^t, t) \in \mathbb{R}$, where $X_i(o_c, t)$ defines a payment to agent i given that the center is in observation state o_c^t at time t.*

Because this is an environment with uncertainty—even on the part of the center—about the current global system state $s \in S$ (due to inaccessibility), the appropriate equilibrium concept is *Bayes-Nash equilibrium*.

Definition 3 (Bayes-Nash Equilibrium). *Given dynamic mechanism $(\boldsymbol{\pi}, X)$ and agents' beliefs $b^t = (b_1^t, ..., b_n^t)$ at time t (where beliefs are updated according to Bayes' rule), strategy profile $\langle f, g \rangle$ constitutes a Bayes-Nash equilibrium if and only if every agent i's expected discounted utility going forward is maximized by following strategies $\langle f_i, g_i \rangle$, given that all other agents $j \neq i$ follow strategies $\langle f_j, g_j \rangle$.*

A mechanism is *Bayes-Nash incentive compatible* if at any time t, for any joint belief state, truthful reporting and taking local actions obediently is a Bayes-Nash equilibrium.

Mechanism 1 (Dynamic-Groves for PS-DEC-MDPs). *At every time step t:*

1. *Each accessible agent can report a claim $f_i(b_i^t, t) = \hat{s}_i^t$ about its current state.*
2. *The center forms o_c^t and computes $\boldsymbol{\pi}^*(o_c^t) = (\pi_c^*, \pi_1^*, \ldots, \pi_n^*)$, the optimal meta-policy given the agent models and current observation state. The center communicates the local policies to all agents that sent a report about state.*
3. *The center executes $\pi_c^*(s_c^t) = a_c$, and each agent executes local action $g_i(b_i^t, t)$.*
4. *The center pays every agent i a transfer:*

$$X_i(o_c^t, t) = \sum_{j \in I \setminus \{i\}} E\big[R_j^t | o_c^t, \boldsymbol{\pi}^*(o_c^t)\big]$$

which is the sum of all other agents' expected rewards. Here, R_j^t is short-hand for agent j's expected reward, where the expectation is taken with regard to the center's current observation state o_c^t, the local policies just prescribed to the agents who reported to be accessible, and the most recent local policy prescribed to the inaccessible agents when they were last accessible.

For simplicity, we ignored the center's local reward $R_c(s_c, a_c)$. We assume it to be zero, but it could be added and the analysis would still go through the same way. In presenting this mechanism, we also assume for now that it is possible to provide a payment to each agent in each period, whether or not the agent is inaccessible. In fact this is precluded, and we explain how to handle this below.[5]

[5]We must also avoid these payments for reasons of computational complexity. If inaccessible agents could receive payments, this would provide a new form of implicit communication. An inaccessible agent could deduce new information about other agents by observing its payments and would have to reason about them.

Lemma 1. *The dynamic-Groves mechanism in which payments are made in every period to every agent is Bayes-Nash incentive compatible in the PS-DEC-MDP model.*

Proof. Consider some period t and assume agents $j \neq i$ follow the equilibrium and are truthful. We assume that it is common knowledge that the center is a social planner, i.e., the meta-decision policy is selected to maximize the agents' expected joint rewards (based on reports). Let $\overline{b}_i^t$ denote the special belief state in which agent i knows the true center's observation state when the agent was last accessible. We establish that the agent cannot benefit by deviating *for any* $\overline{b}_i^t$. This immediately establishes that the agent cannot benefit for *any* beliefs (since it holds pointwise for any realization of the center's observation state.) We now adopt beliefs $\overline{b}_i^t$ and proceed by case analysis on whether agent i is accessible in the current state.

(Case a) Agent i is accessible. We can write down agent i's total expected payoff (reward + payments) from time t forward assuming he is truthful:

$$\rho_i^t(\overline{b}_i^t) = E\big[\sum_{k=t}^{T} \gamma^{k-t} R_i(s_i^k, s_c^k, a_i^k, a_c^k) + \gamma^{k-t} X_i(o_c^k, k) | \overline{b}_i^t\big]$$
$$= E\big[\sum_{k=t}^{T} \sum_{j=1}^{n} \gamma^{k-t} R_j(s_j^k, s_c^k, a_j^k, a_c^k) | \overline{b}_i^t\big],$$

where the expectation is taken with respect to the distribution on system states given belief state $\overline{b}_i^t$ which implies a belief over the center's prescribed polices. We see that the agent receives his own intrinsic reward $R_i(s_i^t, s_c^t, a_i^t, a_c^t)$ along with payments in each period that equal the expected reward to all other agents in that period. Observe that when the agent is truthful the system as a whole implements the optimal PS-DEC-MDP policy and the agent's expected payoff is exactly $V^*(o_c^t)$. For contradiction, assume that the agent can achieve better expected payoff through some non-truthful strategy $\langle f_i', g_i' \rangle$: by the principle of one deviation, consider a deviation in $\langle f_i', g_i' \rangle$ in the current period only.

(1) if agent i misreports his local state to the center the center forms an incorrect observation state. The effect of the deviation is only to change the actions taken by the center and other agents (indirectly via different prescribed policies from the center due to a different observation state). But there exists an equivalent meta-decision policy $\boldsymbol{\pi}'$ that would implement the same actions, and thus this is a contradiction because it implies that the center's current meta-decision policy is suboptimal (given that agent i's payoff is aligned with the total system reward, and the agent's beliefs $\overline{b}_i^t$ are those of the center.)

(2) if agent i deviates from the prescribed local strategy this affects his immediate local reward, his local state transition and, depending on future state reports, provides the center with a different observation state. Agent i could affect (and indeed increase his local reward) directly, but only affect his payment (other agents' rewards) indirectly via new policies issued by the center. This is true because of the specific independence assumptions made in the PS-DEC-MDP model. Again, there exists an equivalent meta-decision policy $\boldsymbol{\pi}'$ that would implement the same actions and this is a contradiction with the optimality of the current meta-decision policy.

It is also easy to see from the same argument that no combination of deviations of kinds (1) and (2) is useful.

(Case b) Agent i is inaccessible. The argument here is essentially the same. Consider a deviation $\langle f_i', g_i' \rangle$ in the current period only. The only deviation to consider now is one in selecting the agent's local action. When choosing his local action the agent will compute his expected future payoff given his belief state $\overline{b}_i^t$ and taking into consideration that he will only be able to communicate with the center again and make a state report at some point in the future. But again, the agent is computing the expectation with respect to the same information available to the center when he computed the meta-decision policy; in this case, this is the center's observation state $o_c^{t'}$ in period t' when the agent was last accessible. In particular, if there is a useful action deviation for agent i that would increase his expected payoff this would contradict our assumption that $\boldsymbol{\pi}^*$ was optimal because the center could have "programmed" this deviation into the emergency policy provided to agent i and triggered this deviation in the agent's current local state. □

We have shown that even if agent i has enough information to reason about the center's observation state correctly, he doesn't need to do so, because for any possible beliefs he is best off following the prescribed local policy (given that the center is acting as a social planner and in equilibrium with the other agents acting truthfully).

The proof relied heavily on the specific independence assumptions made in the PS-DEC-MDP model. It is important that a misreport or an action deviation by agent i only indirectly affects the other agents via future changes in the policies assigned by the center. If that were not the case, the mechanism would not be Bayes-Nash incentive compatible. We can construct a simple counter-example: Assume we have a PS-DEC-MDP with two agents and one center. Agent 1 has only one local state and one action but agent 2 has one local state and two actions a_1 and a_2. Agent i's reward function is dependent on agent 2's local action such that he receives a very high reward when a_1 is chosen but a very low reward when a_2 is chosen. Agent 2's reward function is set up such that a_1 leads to a small and a_2 only to a slightly higher reward. Thus, the efficient meta-decision policy that maximizes the joint reward would assign action a_1 to agent 2 and the Dynamic-Groves mechanism would pay agent 2 the expected reward of agent 1 given the meta-decision policy. However, because payments are made based on expectations and not on realized rewards, agent 2 maximizes his payoff by taking action a_2, getting the slightly higher local reward and still being paid the high reward that was expected (but not received) for agent 1.

To modify the mechanism to handle the constraint that payments cannot be made while an agent is inaccessible we can simply roll-up the payments that inaccessible agents would otherwise receive and make a "lump sum" payment to such an agent as soon as it becomes accessible. This, however, requires the additional assumption that *each agent must*

eventually make any payments it owes to the center. This is required to preclude that an agent who still owes payments to the center would choose to "hide for ever".

With $X(o_c^t, i)$ still used to denote the payment made in the dynamic-Groves mechanism (as described above) to an agent at time t, the payment scheme in step (4.) of the modified mechanism is:

The center pays every agent i that makes a report about his state:

$$\hat{X}_i(o_c^t, t) = \sum_{k=t-\delta(t)}^{t} \frac{X_i(o_c^k, k)}{\gamma^{t-k}},$$

where $\delta(t) \geq 0$ is the number of successive periods prior to t that i has been inaccessible and γ is the discount factor.

This reduces to the same payment term when an agent is accessible, while ensuring that the expected payment of an inaccessible agent forward from any state is equal to that in the earlier mechanism; see also Cavallo et al. (2007). This is easy to see when one observes that the payment $X_i(o_c^t, t)$ is already independent of agent i's strategy until the agent is again accessible because the center's observation state in regard to agent i does not change until it receives a new report about local state and the agent's actions do not influence any other agent's states. We refer to the modified mechanism as the dynamic-Groves# mechanism.

Theorem 6. *The dynamic-Groves# mechanism is efficient and Bayes-Nash incentive compatible in the PS-DEC-MDP model.*

Conclusion

In this paper we presented the first work that combines models and techniques from decentralized decision making under uncertainty with dynamic mechanism design. We specifically addressed environments in which agents periodically become inaccessible, yet can still take local actions and undergo state transitions—for instance, when a group of self-interested taxi drivers might occasionally lose communication with the dispatcher, or in Internet settings where a set of network servers must be coordinated and communication links are fragile or faulty. In formalizing these domain characteristics, we introduced the *partially-synchronized DEC-MDP* framework that models precisely the specific independence and communication limitations required for positive complexity results and successful incentive design.

In considering incentive issues related to private actions, the main challenge was that local actions are unobservable and thus not "contractible," so that the center's decision policy must also be individually optimal for every agent in equilibrium. As we illustrated in the counter-example, it is necessary that local agents' rewards are independent of each other's actions, even when they are accessible, which is a requirement over-and-above that which would be required just for computational tractability.

We proved a series of complexity results for the PS-DEC-MDP model including on the negative side that the infinite-horizon problem with no limits on inaccessibility is undecidable, and on the positive side that the finite and the infinite-horizon problem for a fixed number of agents with a priori limits on inaccessibility is P-complete. Finally, we presented "Dynamic Groves#", an efficient dynamic mechanism for managing incentives in a context of self-interested agents.

For future research we are considering ways to relax the strong independence assumptions in our model using the techniques introduced in Mezzetti (2004). Another interesting next step is the extension of this work to online scenarios with changing agent populations where each agent has a fixed arrival and departure time. We hope that this paper will inspire development of computationally efficient algorithms required for the PS-DEC-MDP model, together with other directions in bridging the fields of dynamic mechanism design and decentralized decision making under uncertainty.

Acknowledgments

We thank Marek Petrik and Avi Pfeffer for very helpful discussions on this work.

References

Athey, S., and Segal, I. 2007. An Efficient Dynamic Mechanism. Working Paper, Department of Economics, Harvard University.

Becker, R.; Zilberstein, S.; Lesser, V.; and Goldman, C. V. 2004. Solving transition independent decentralized Markov decision processes. *Journal of Artificial Intelligence Research (JAIR)* 22:423–255.

Bergemann, D., and & Välimäki, J. 2006. Efficient dynamic auctions. Cowles Foundation Discussion Papers 1584.

Bernstein, D. S.; Givan, R.; Immerman, N.; and Zilberstein, S. 2002. The complexity of decentralized control of Markov decision processes. *Mathematics of Operations Research* 27(4):819–840.

Boutilier, C. 1999. Sequential optimality and coordination in multiagent systems. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI)*, 478–485.

Cavallo, R.; Parkes, D. C.; and Singh, S. 2006. Optimal coordinated planning amongst self-interested agents with private state. In *Proceedings of the Twenty-second Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 55–62.

Cavallo, R.; Parkes, D. C.; and Singh, S. 2007. Efficient online mechanisms for persistent, periodically inaccessible self-interested agents. In *DIMACS Workshop on the Boundary between Economic Theory and Computer Science*.

Goldman, C. V., and Zilberstein, S. 2004. Decentralized control of cooperative systems: Categorization and complexity analysis. *Journal of Artificial Intelligence Research (JAIR)* 22:143–174.

Madani, O. 2000. *Complexity results for infinite-horizon Markov decision processes*. Ph.D. Dissertation, University of Washington.

Mezzetti, C. 2004. Mechanism design with interdependent valuations: Efficiency. *Econometrica* 72(5):1617–1626.

Papadimitriou, C. H., and Tsitsiklis, J. N. 1986. Intractable problems in control theory. *SIAM Journal on Control and Optimization* 24(4):639–654.

Parkes, D. C. 2007. Online Mechanisms. In *Algorithmic Game Theory*, Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay Vazirani (eds.) , Chapter 16, Cambridge University Press.

Seuken, S., and Zilberstein, S. 2008. Formal models and algorithms for decentralized decision making under uncertainty. *Autonomous Agents and Multi-Agent Systems*. DOI: 10.1007/s10458-007-9026-5.