Efficient Online Mechanisms for Persistent, Periodically Inaccessible Self-Interested Agents

Ruggiero Cavallo
SEAS, Harvard University
cavallo@eecs.harvard.edu

David C. Parkes
SEAS, Harvard University
parkes@eecs.harvard.edu

Satinder Singh
Computer Science and Engineering
University of Michigan
baveja@umich.edu

First draft: June 5, 2007
This draft: July 19, 2007

Abstract

We consider the problem of implementing a system-optimal decision policy in the context of self-interested agents with private state in an uncertain world. Unique to our model is that we allow both *persistent* agents, with an agent having a local MDP model to describe how its local world evolves given actions by a center, and also *periodically-inaccessible* agents, with an agent unable to report information while inaccessible. We first review the dynamic-VCG mechanism of Bergemann and Välimäki (2006), which handles persistent agents. We offer an independent, simple proof of its correctness. We propose a generalized mechanism to allow also for inaccessibility, and identify conditions for its correctness. In closing, we observe that the mechanism is equivalent to the earlier online-VCG mechanism of Parkes and Singh (2003) in a restricted setting.

1 Introduction

Mechanism design (MD) is the problem of "inverse game theory." The setting is one of multiple self-interested agents, each with private inputs that are relevant to a decision and with a utility function on decisions. The problem is to design a game such that, in the non-cooperative equilibrium, the decision selected implied by the outcome of the game satisfies some desired set of properties. The Vickrey-Clarke-Groves (VCG) mechanism (see, e.g., [Jackson, 2003]) is a celebrated solution that provides (economic) *efficiency*, i.e., a decision that maximizes the total utility of agents, in a simple, dominant-strategy equilibrium. The VCG mechanism also runs without a deficit in reasonable environments, so that the center does not need to subsidize the incentive mechanism.

In extending MD to dynamic environments, and while retaining the goal of efficiency, one seeks to implement a *sequence* of utility-maximizing decisions in an uncertain environment. Agents' private inputs provide information about state, reward, available actions, and dynamics. Two kinds of problem variants have been studied in the literature. In one variant, the agents are *persistent* and the agent population fixed, while each agent receives private information over time, perhaps in a way that depends on decisions that are made by the center [Cavallo *et al.*, 2006; Bergemann and Valimaki, 2006]. In another variant, the agent population is dynamic, with each agent *inaccessible* for some period of time, but once accessible an agent knows all of its private information and can report it in a single period [Lavi and Nisan, 2000; Parkes and Singh, 2003]. An inaccessible agent cannot send messages to the center and cannot be charged.[1]

Unique to our model is that we allow both persistent and periodically-inaccessible agents. We first review the dynamic-VCG mechanism of Bergemann and Välimäki (2006), which handles persistent agents. We offer an independent, simple proof of its correctness. We propose a generalized mechanism, *dynamic-VCG#*, to allow for periods of inaccessibility together with stochastic local dynamics, and identify conditions for its correctness. In doing so, we are able to unify these two threads of research and significantly expand the domains to which dynamic mechanisms can be applied. We observe that the second mechanism is equivalent to the earlier online-VCG mechanism of Parkes and Singh (2003) in a restricted model. We close with some remarks to indicate the breadth of multi-agent domains that can be coordinated via these mechanisms.

2 A Fixed Population of Accessible Agents

First consider a standard multi-agent environment, with a fixed set of $N = \{1, \ldots, n\}$ agents able to communicate with a central decision maker (center). Each agent i has a private and local state ($\in S_i$) that evolves over time depending on the decisions taken by the center. The center has state $s_0 \in S_0$, which collects additional information to make this an MDP. For example, state $s_0 \in S_0$ can be used to keep track of actions. We denote the joint state space by $S = S_0 \times S_1 \times \ldots \times S_n$ and the state space with i hidden as S_{-i}. The set of decisions is A and the center chooses from feasible decisions $A(s) \subseteq A$ in each state $s \in S$, over a time horizon of K (which may be infinite). The dynamics for agent i are defined by a stochastic transition function $\tau_i : S \times A \to S_i$ such that for all $s \in S$ and $a \in A$,

[1] We refer the interested reader to Parkes (2007) for a survey of online MD. Athey and Segal (2007) also work in the persistent, accessible model and provide an interim incentive-compatible mechanism that is budget-balanced on average.

$\sum_{s'_i \in S_i} P(\tau_i(s,a) = s'_i) = 1$ (for prob. function P). Similarly, agent i receives reward $r_i(s,a)$ when the center takes action a in joint state s. Thus, agent i is defined by a time-invariant MDP model $M_i =< S_i, A, \tau_i, r_i >$. Both the model and the agent's state in any period are private to the agent. It is convenient to include the agent's model as part of its $t = 0$ state.

The goal of the center is to maximize the discounted summed rewards obtained by the agents over the time horizon K. Let s^t, s_i^t, and s_{-i}^t denote respectively the joint state, agent i's state, and the joint state of all agents but i at time t. Furthermore, let π be a decision policy that maps joint states to actions. We define $V_i^\pi(s)$ to be agent i's expected value for π given state s, i.e., $V_i^\pi(s) = \mathbb{E}_{s^{\geq t}}[\sum_{k=t}^K \gamma^{k-t} r_i^k(s^k, \pi(s^k))]$, where the expectation is taken w.r.t. the distribution on future states, denoted $s^{\geq t} = (s^t, \ldots, s^K)$, with $s^k = \tau(s^{k-1}, \pi(s^{k-1})), \forall k > t$, and where $0 < \gamma \leq 1$ is the discount factor. We write $r(s,a)$ to denote $\sum_{i \in N} r_i(s,a)$, $V^\pi(s)$ to denote $\sum_{i\in N} V_i^\pi(s)$, and $V_{-i}^\pi(s)$ to denote $\sum_{j \in N \setminus \{i\}} V_j^\pi(s)$. We use π^* to denote the optimal decision policy (in space of all decision policies Π), i.e., $\pi^* = \arg\max_{\pi \in \Pi} V^\pi(s), \forall s \in S$. We write $V^*(s)$ as shorthand for $V^{\pi^*}(s)$. We will at times consider the policy that is optimal over a subset of agents; π_{-i}^* will denote the policy optimal for $N \setminus i$, i.e., $\pi_{-i}^* = \arg\max_{\pi_{-i} \in \Pi_{-i}} V_{-i}^{\pi_{-i}}(s), \forall s \in S$. We write $V_{-i}^{\pi_{-i}}(s)$ rather than $V_{-i}^{\pi_{-i}}(s_{-i})$ because agent i remains *present* in the world even though its value is ignored. This distinction can matter when there is an interdependence between agents; e.g., with the state s_i influencing the reward of some agent $j \neq i$. For convenience we adopt notation $V_{-i}^*(s_{-i})$ for $V_{-i}^{\pi_{-i}^*}(s)$, where agent i's state is left implicit, because we will make an independence assumption that removes this issue except for dynamic populations (in Section 3.2).

2.1 Online Mechanisms

An *online mechanism* is defined by a decision policy π and a payment policy T, which maps reported state information to a payment made **to** each agent (note the sign convention). Formally, $T = \{T_1, \ldots, T_n\}$, and $\forall i \in N$, $T_i : S \rightarrow \mathcal{R}$. Each agent i will report state information according to some *strategy* $f_i : S_i \rightarrow S_i$. We use F_i to denote the set of all strategies available to agent i (i.e., all possible mappings of a true state to a reported state).[2] Note that report $f_i(s_i^0)$ includes the agent's model in period $t = 0$.[3] We write $f(s)$ to denote $(s_0, f_1(s_1), \ldots, f_n(s))$, i.e., the reported joint state when the true joint state is s. Hereafter, a policy π is a mapping from *reported state* to action because the center's view of state s is limited to $f(s)$. Fix some policy π. Let $\mathbb{E}_{s^{\geq t}}[\sum_{k=t}^K \gamma^{k-t} g(s^k) | f_i] = \mathbb{E}_{\mathfrak{s}^{\geq t}}[\sum_{k=t}^K \gamma^{k-t} g(\mathfrak{s}^k)]$, where $\mathfrak{s}^k$ is the state reached in period k given that agent i misreports its local state according to

[2] We can assume that any strategy for agent i depends only on the *current* state, as any historical state or decision information can be incorporated into the current state representation.

[3] For simplicity, we assume that an agent cannot make a misreport that materially changes the set of available actions that the center believes are available. Such a misreport could be caught and punished with a large fine.

f_i and the rest are truthful. This expectation is, throughout the paper, taken w.r.t. the true joint model. We assume quasilinear utility, so that net utility in period t is the expected discounted reward *plus* expected discounted payments.

Definition 1 (interim incentive compatibility). *A dynamic mechanism* (π, T) *is interim incentive compatible if and only if, at all times t, for all agents i, for all possible true states s^t, and for all f_i,*

$$\mathbb{E}_{s^{\geq t}}\Big[\sum_{k=t}^{K}\gamma^{k-t}\Big(r_i(s_i^k, \pi(s^k)) + T_i(s^k)\Big)\Big] \geq \mathbb{E}_{s^{\geq t}}\Big[\sum_{k=t}^{K}\gamma^{k-t}\Big(r_i(s_i^k, \pi(f_i(s_i^k), s_{-i}^k)) + T_i(f_i(s_i^k), s_{-i}^k)\Big) \,\Big|\, f_i\Big] \quad (1)$$

A mechanism is *interim incentive compatible* (IC) if each agent maximizes his *payoff* (or expected, discounted utility) by reporting truthfully, given that the other agents do the same. This includes truthfully reporting its model in period $t = 0$.[4]

The following impossibility result from static MD motivates an additional requirement that we impose in our dynamic environment.

Proposition 1 (entailed by [Jehiel and Moldovanu, 2001], Theorem 4.3). *In static environments where agent valuations may be arbitrarily interdependent, there exists no efficient*[5] *and ex post incentive compatible mechanism.*

Interdependent valuations are those in which one agent's utility for a decision depends on the private (valuation) information of another agent. Without a further restriction, we can provide a reduction from the static, interdependent value problem to the dynamic, multi-agent model.

Theorem 1. *In arbitrary time-horizon dynamic environments where agent reward functions can arbitrarily depend on other agents' states, no mechanism is efficient and interim incentive compatible.*

Proof. The theorem follows immediately from Proposition 1, as any mechanism that is interim IC for arbitrary time-horizon dynamic settings must be so for single time-step (static) settings. Proposition 1 tells us that when agent values (here, $r_1, \ldots, r_n$) can depend arbitrarily on other agents' types (here, $s_1^0, \ldots, s_n^0$), no such mechanism exists. □

Theorem 2. *In arbitrary time-horizon dynamic environments where agent transition functions can arbitrarily depend on other agents' states, no mechanism is efficient and interim incentive compatible.*

[4] Note that it does not matter whether or not the agent *knows* the current joint state s^t or the joint transition model, because the inequality is established for all possible current joint states and all possible joint models, under the assumption that the other agents report truthfully.

[5] We will use the term "efficient" for any mechanism that achieves a decision policy that maximizes utility summed over all agents.

Proof. Again, the theorem follows from Proposition 1. Assume for contradiction the existence of a dynamic mechanism $M = (\pi^M, T^M)$ that is interim IC for arbitrary time-horizon dynamic settings, where agent transition functions can arbitrarily depend on other agents' states. Now consider an arbitrary static decision problem with a set of possible outcomes O and agent private signals $\theta_1, \ldots, \theta_n$ and valuations $v_1, \ldots, v_n$, where each v_i is a function of i's private signal θ_i and the other agents' valuations v_{-i}. If the theorem holds, then one can construct an interim IC 2 time-step dynamic mechanism that corresponds to an interim IC mechanism for the static case. In the first time-step, each agent transitions to a state that incorporates its own private signal report and the other agents' valuations, conditioned on the agent's signal report. The second time-step has transitions to a single "final-state" with rewards that correspond to the agent's value, which has been conditioned on the previous state. Applying mechanism M to a dynamic world constructed this way would require exactly one report of private information (just as would be required in the static case), and would constitute a solution to the static problem. A demonstration with a 2-agent, 2-outcome example is provided in Figure 2 in the appendix. □

In light of these theorems we will restrict our attention to environments without informational externalities; for the rest of the paper we make the following assumption:[6]

Assumption A 1. *Each agent's reward and transition functions are conditionally independent of other agents' states given an action, i.e.,* $\forall i \in N; \forall s_i \in S_i; \forall s_{-i}, s'_{-i} \in S_{-i}; \forall a \in A$, *we have* $r_i((s_i, s_{-i}), a) = r_i((s_i, s'_{-i}), a)$ *and* $\tau_i((s_i, s_{-i}), a) = \tau_i((s_i, s'_{-i}), a)$.

We will accordingly write $r_i(s_i, a)$ and $\tau(s_i, a)$ to denote, respectively, an agent's reward and transition when action a is taken while i is in state s_i, regardless of s_{-i}.

2.2 The Dynamic-VCG Mechanism

The dynamic-VCG mechanism makes decisions according to the optimal policy (given reported models and applied to reported state), and specifies, at every time step, a payment **to** each agent i equal to i's "flow marginal contribution" at that time-step, i.e., the positive impact that i has on the ability for the *other* agents to obtain value in the current time-step and in the future. This impact is via i's presence in the world, its model, and its current state and occurs indirectly, through the impact of i on the decisions made by the policy. In defining the dynamic-VCG mechanism, we write $Q^*(s, a) = r(s, a) + \gamma \mathbb{E}_s[V^*(\tau(s, a))]$, and the same "${}_{-i}$" syntax for its variants without agent i as was defined for V^*.

[6] An assumption this strong is technically not required to achieve efficiency and interim incentive compatibility, but in this paper we do not wish to delve into technical requirements akin to the single-crossing condition (see [Cremer and McLean, 1985]), so we use this broad stroke.

Mechanism 1 (Dynamic-VCG). *[Bergemann and Valimaki, 2006]*

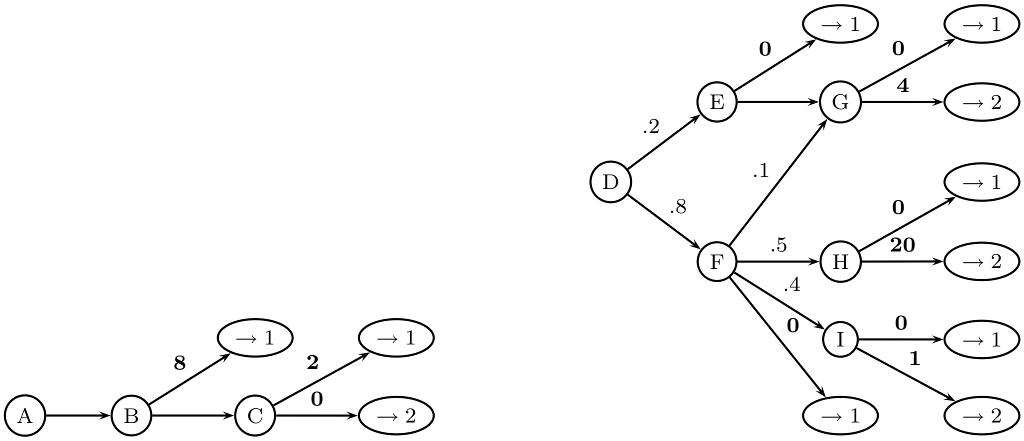
At every time step t (in true state s^t):

1. *Each agent i reports to the center a claim, $f_i(s_i^t)$, about its current state.*
2. *The center executes action $\pi^*(f(s^t))$, where π^* is optimal given reported agent models.*
3. *The center pays to each agent a payment:*

$$\begin{aligned} T_i^t(f(s^t)) &= V_{-i}^*(f_{-i}(s_{-i}^t) \,|\, \pi^*(f(s^t))) - V_{-i}^*(f_{-i}(s_{-i}^t)) \\ &= Q_{-i}^*(f_{-i}(s_{-i}^t), \pi^*(f(s^t))) - V_{-i}^*(f_{-i}(s_{-i}^t)), \end{aligned} \tag{2}$$

where the expected values (V_{-i}^*, Q_{-i}^*) are taken w.r.t. the reported agent models.

Note that in the first period only, part of an agent's report is a claim about its MDP model. Agents make claims about states directly, but claims about rewards indirectly via the model described in $t = 0$. The payment to agent i in dynamic-VCG is equal to the difference between the value the other agents get from the action selected because agent i is present, followed by the optimal sequence of actions for agents $\neq i$ in the future, and the value they would get from the optimal sequence of actions forward from the current state.



(a) Agent 1's MDP.

(b) Agent 2's MDP.

Figure 1: Two-agent, 3 time-step world. Actions ({allocate to 1, allocate to 2, don't allocate}) are implicit in the state transitions. Agent 1's MDP has deterministic transitions, while agent 2's has uncertain transitions in the first one or two time-steps.

Example 1. Consider the simple two agent example portrayed in Figure 1.[7] Assume discount factor $\gamma = 1$ (i.e., no discounting). The optimal policy allocates to agent 1 in state BE, to agent 2 in states $\{CG, CH\}$, to agent 1 in state CI and makes no allocation in states $\{AD, BF\}$. Because of the special structure of this domain, the VCG payment to agent i in state s^t is $-V^*_{-i}(s^t_{-i})$ when it is allocated, because $Q^*_{-i}(s^t_{-i}, \pi^*(s^t)) = 0$ since the other agent cannot get the item. Otherwise, the payments are always 0 except when the presence of agent i in state s^t_i precludes the other agent from being allocated now. In this case, the payment is the cost (if any) of a delay in this decision. To be concrete, consider (true) state BE. If agent 2 reports E ("med value") agent 1 is allocated for payment -4 and agent 2's payoff is 0. If agent 2 reports F ("poss. high value") its payment is -6 (the externality it imposes on agent 1.) Continuing, the (true) next state is CG. If agent 2 reports G or H its payment is -2, if it reports I agent 1's payment is -1. Agent 2's best misreport is G or H, but its net payoff from either deviation is $-6 + 4 - 2 = -4$, so it should have reported state E truthfully. Other misreports can be checked, and none are useful. The up-shot is that agent 2 will truthfully report states E and I when they occur, and the center gains the information it needs to know when to allocate to agent 1.

2.3 Dynamic-VCG is Interim Incentive Compatible

We proceed by offering a simple proof for the correctness of the dynamic-VCG mechanism in an environment with persistent, accessible agents. Our proof is short, and emphasizes the connection to the simple theory of (static) *Groves* mechanisms in which incentives are aligned via a transfer equal to the reward to every other agent, coupled with some other, agent-independent term.[8] Let $V^\pi(s^t|f_i) = \mathbb{E}_{s \geq t}\Big[\sum_{k=t}^{K} \gamma^{k-t} r(s^k, \pi(f_i(s^k_i), s^k_{-i}))|f_i\Big]$ denote the total expected discounted reward forward from state s^t, given policy π, where the expectation is taken w.r.t. the *true* joint model, and with agent i adopting strategy f_i in misreporting its state.

Lemma 1. *A dynamic mechanism* (π, T) *is interim incentive compatible with persistent, always-accessible agents, if:*

i) $\forall s \in S$, *policy* $\pi(s) = \pi^*(s)$, *where policy* π^* *is optimal given reported agent models, and*

ii) agent i*'s expected payoff, (with respect to the true joint model) in any true state* s^t, *given strategy* f_i, *and given that all other agents report truthfully, is:*

[7] Nodes represent states, the initial joint state is AD, probabilistic transitions are annotated with the probability $(.x)$. The terminal states are denoted → 1 or → 2 to indicate a joint action a was taken that allocated to agent 1 or 2, respectively. Only these actions have non-null rewards, and these rewards are indicated in **bold**.

[8] Bergemann and Välimäki (2006), who discovered this mechanism, provide an alternate proof. Cavallo et al. (2006) earlier proposed a related mechanism, but it satisfies the weaker property of ex ante IR, meaning that agents must commit to the mechanism at time $t = 0$.

$V^{\pi}(s^t|f_i) - C_i(s^t)$, where $C_i(s^t)$ is a constant and, given we are in state s^t, independent of strategy f_i (i.e., independent of the reported model, and also current and future reported states).

Proof. Fix agent i and suppose agents $\neq i$ are truthful. Assume, for contradiction, that IC fails. Then, there must be some strategy f_i and some state s^t, for which

$$V^{\pi}(s^t|f_i) - C_i(s^t) > V^*(s^t) - C_i(s^t), \tag{3}$$

where the form of the LHS and RHS follow from property (ii), and the RHS is the payoff to agent i from reporting truthfully, by property (i). But now, if $V^{\pi}(s^t|f_i) > V^*(s^t)$ for misreport f_i, where the model it reports influences the choice of π and its state misreports influence the way in which π is applied, then we can construct policy $\pi'(s^k) = \pi(f_i(s^k_i), s^k_{-i})$ on the underlying (true) states with $V^{\pi'}(s^t) > V^*(s^t)$, which is impossible by the definition of an optimal MDP value function. □

Each agent's payoff is aligned with that of the total value achieved by the system given policy π and strategy f_i, which is maximized by truthful reports so that the policy is optimal. Agent i's payoff is affected by $C_i(s^t)$, but this is conditionally independent of the agent's strategy given the current state s^t.

Theorem 3. *The dynamic-VCG mechanism is interim incentive compatible (at every time step) with persistent, always-accessible agents.*

Proof. Property (i) in Lemma 1 holds by construction. Fix some agent i, strategy f_i, some state s^t, and assume agents $\neq i$ are truthful. Given policy π^*, optimal w.r.t. the reported model of agent i and true model of the other agents, the payoff to agent i state is:

$$\mathbb{E}_{s\geq t}\Big[\sum_{k=t}^{K}\gamma^{k-t} r_i(s^k_i, \pi^*(f_i(s^k_i), s^k_{-i})) + \sum_{k=t}^{K}\gamma^{k-t}(Q^*_{-i}(s^k_{-i}, \pi^*(f_i(s^k_i), s^k_{-i})) - V^*_{-i}(s^k_{-i}))\;\Big|\; f_i\Big] \tag{4}$$

The expectation is taken w.r.t. the true joint model, for states reached given that agent i plays f_i, with the Q^*_{-i} and V^*_{-i} terms defined for the optimal policy without agent i and w.r.t. the correct model of agents $\neq i$ (since they are truthful). This is equivalent to:

$$V^*(s^t|f_i) - \mathbb{E}_{s\geq t}\Big[\sum_{k=t}^{K}\gamma^{k-t}(V^*_{-i}(s^k_{-i}) - \gamma V^*_{-i}(s^{k+1}_{-i}))\;\Big|\; f_i\Big], \tag{5}$$

where $V^*(s^t|f_i)$ comes from combining the first term in Eq. (4) with the stream of single-period rewards to the other agents within the Q^*_{-i} term in Eq. (4) (and leveraging assumption A1, by which $\sum_{j\neq i} r_j(s,a) = \sum_{j\neq i} r_j(s_{-i}, a)$.) In the summation, the expected values of component $V^*_{-i}(s^k_{-i})$ (directly from Eq. (4)) and component $-\gamma V^*_{-i}(s^{k+1}_{-i})$ (the continuation value for expanding the corresponding $Q^*_{-i}(s^k_{-i})$ term in Eq. (4) for one period) are both taken w.r.t. the same distribution, i.e. on states distributed according to policy π^* on the joint state given strategy f_i. Now consider the second term in Eq. (5); this equals:

$$-V^*_{-i}(s^t_{-i}) - \mathbb{E}_{s^{\geq t}}\Big[\sum_{k=t+1}^{K}\gamma^{k-t}V^*_{-i}(s^k_{-i}) - \gamma\sum_{k=t}^{K}\gamma^{k-t}V^*_{-i}(s^{k+1}_{-i})\,\Big|\,f_i\Big] = -V^*_{-i}(s^t_{-i}), \qquad (6)$$

where Eq. (6) follows since $V^*_{-i}(s^{K+1}_{-i}) = 0$, and by a simple change of variable in the index of the second summation. This completes the proof by appeal to Lemma 1 since $V^*_{-i}(s^t_{-i}) = V^{\pi^*_{-i}}_{-i}(s^t)$. Conditioned on state s^t, this value is independent of the reported model and f_i because of the independence across agents provided by Assumption A1. □

We see that the dynamic-VCG mechanism is defined so that each agent i's expected discounted utility in equilibrium forward from any state s^t is

$$V^*(s^t) - V^*_{-i}(s^t_{-i}) \qquad (7)$$

Given this, and with the reasonable assumption of *non-negative marginal product* (NNMP)[9] such that $V^*(s^t) \geq V^*_{-i}(s^t_{-i})$ in each period, we have:

Theorem 4. *The dynamic-VCG mechanism is interim individual rational at every time-step with persistent, always-accessible agents and non-negative marginal product.*

Interim-individual rationalty means that it is rational for every agent to continue in every period; expected payoff will be non-negative for doing so. On the other hand, the dynamic-VCG mechanism is not *ex post* individual rational. Consider again the example in Figure 1. Agent i's payment in state F is -6, but if it transitions to state G or I in the next period, its final payoff is $-6+4-2=-4$ or -6 respectively. On the other hand, its *expected* payoff is non-negative forward from every state. In state F, for example, its expected payoff is $p(G|F)(-4) + p(H|F)(12) + p(I|F)(-6) = (0.1)(-4) + (0.5)(12) + (0.4)(-6) = 3.2$. Thus non-negative payoff is achieved in expectation, and not ex post.

3 Introducing Periods of Inaccessibility

We consider now the additional possibility that an agent may be inaccessible for some period of time. By inaccessible, we mean that an agent *cannot report any information about its local state or be charged by the center.* An agent cannot claim to be accessible (by sending a message) when it is actually inaccessible, but can

[9] This would be expected to hold unless an agent, just by its presence, negatively effects the total value that is possible in the system (including the value to itself). A setting with physical congestion might violate this; e.g., an extra robot prevents any robot from doing anything useful.

pretend to be inaccessible when it is in fact accessible. We model a "report" of inaccessibility with the *null* message $f_i(s_i^t) = \phi$. Sections 3.1 and 3.2 consider two different variations.

3.1 Persistent Agents with Periodic Inaccessibility

We continue to consider a fixed set of agents. Each agent i may now also be accessible or inaccessible to the center. We have in mind environments in which an agent might lose communication with the center, or leave and do something else, for a while. States now include whether agents are accessible, and $H(s) \subseteq N$ defines the set of accessible ("here") agents. For simplicity, we assume every agent is accessible and able to report a model at $t = 0$. Our results allow transitions and rewards to depend on actions while an agent is inaccessible, and agent accessibility can depend on previous actions. The main question we ask is the following: *Can we design an efficient mechanism in which an agent will truthfully report its state information whenever it can, i.e., whenever it is accessible?*

To see the new difficulty, consider a simple Groves-based mechanism with a naive policy that ignores the existence of any inaccessible agents, following the optimal policy for just the accessible agents. Couple this with a payment scheme that pays each accessible agent in a period the reward of the other agents based on the action and their reported models.

Example 2. Consider the example in Figure 1, modified so that agent 1 is always accessible and agent 2 is inaccessible in period 0, but will become accessible in period 1 or 2 or, with a negligible probability $\epsilon > 0$, not at all. If agent 2 is not accessible in period 1 then agent 1 should pretend to be inaccessible, to avoid receiving the item and so that agent 2 will receive the item, and likely a higher reward (and thus payment to agent 1) in period 2.

In this environment, the optimal policy should reason about the distribution of possible states for an agent that is currently inaccessible. To model this we adopt the Partially Observable MDP (POMDP) formalism, because while an agent is inaccessible the center may only have partial information about the state of agent i. We formulate this as a *belief-state MDP* [Kaelbling *et al.*, 1996]. Let $BS = S_0 \times BS_1 \times \ldots \times BS_n$, and $BS = \Delta(S_i)$, such that $bs_i \in BS_i$, $i \in N$, defines a probability distribution on agent i's state and bs_0 is used by the center to keep appropriate history, as before. The POMDP transition model is defined so that $bs_i^t = s_i^t$ if $i \in H(s_i^t)$, and updated according to the agent's model and the action taken otherwise.[10] Agent MDPs induce reward $r(bs^t, a) = \sum_i r_i(bs_i^t, a)$ by expectation over the belief state. The optimal policy, $\pi^* : BS \rightarrow A$, maximizes the expected discounted reward in every belief state. The dynamic-VCG mechanism is now defined on belief states:

[10] To avoid conditioning beliefs on the availability of actions, we assume the feasible joint actions depend only on states of accessible agents, and also that agents don't misreport in a way that leads to a misrepresentation of the available actions (the center could catch such a deviation anyway).

Mechanism 2 (Dynamic-VCG). *At every time step t (in state s^t):*

1. *Each accessible agent can report a claim $f_i(s_i^t)$ about its current state.*
2. *The center updates its belief state bs^t, and selects joint action $a^t = \pi^*(bs^t)$, where π^* is the optimal policy given reported agent models.*
3. *The center pays each agent i that makes a report:*

$$T_i^t(bs^t) = Q_{-i}^*(bs_{-i}^t, \pi^*(bs^t)) - V_{-i}^*(bs_{-i}^t)$$

But this mechanism fails.

Example 3. Consider the example in Figure 1 modified so that agent 1 is always accessible, and agent 2 is inaccessible in period 0 but will become accessible in period 1 or 2 or, with a negligible probability $\epsilon > 0$, not at all. If agent 2 is accessible in period 1 and in state E it will claim to be inaccessible. Why? If truthful, agent 1 is allocated and agent 2's payoff is zero. By lying, the policy will delay making an allocation until period 2 because $8 < (0.2)4 + (0.8)((0.1)4 + (0.5)20 + (0.4)2) = 9.76$ (ignoring ϵ). Both agents' payments in period 1 will be zero (agent 2's because it is inaccessible). Agent 2 can now report state G in period 2 and receive the item, for a payment of -2 and a net payoff of $4 - 2 = 2$. Note the efficiency loss: the planner should have allocated to agent 1 in period 1.

Dynamic-VCG satisfies the corresponding notion of property (i) in Lemma 1 for this environment, but fails to satisfy property (ii). To understand this, define a *true belief state*, bs^t, as the belief state the center would be in, given some policy π, if *every* agent is truthful and reports its state whenever it is accessible. Dynamic mechanism (π, T) is IC in this environment, if for any agent i, with agents $j \neq i$ truthful, and in any true belief state bs^t, agent i maximizes its payoff by following the truthful strategy. A corresponding version of Lemma 1 for this environment is:

Lemma 2. *A dynamic mechanism (π, T) is interim incentive compatible with persistent, periodically-inaccessible agents, if:*

i) policy π is optimal given reported models, and

ii) agent i's expected payoff (w.r.t. the true model), in any true belief state bs^t, given strategy f_i, and given that the other agents are truthful, is $V^\pi(bs^t|f_i) - C_i(bs^t)$, where $C_i(bs^t)$ is a constant, and independent of strategy f_i.

Proof. Fix agent i and agents $j \neq i$ to be truthful. Assume IC fails. Then there must be some f_i and some true belief state bs^t, for which $V^\pi(bs^t|f_i) - C_i(bs^t) > V^*(bs^t) - C_i(bs^t)$. But, we can then construct an equivalent policy $\pi'(bs^k) - \pi(bs^k|f_i)$, where $\pi(bs^k|f_i)$ is policy π applied to the belief state the center would have if agent i

had followed f_i rather than being truthful. But now $V^{\pi'}(bs^t) > V^*(bs^t)$, and a contradiction. □

The dynamic-VCG mechanism cannot achieve this in general because the payments cannot be made in periods during which the agent is inaccessible, and thus its payoff is not correctly aligned. In the example, the agent is able to mimic the effect of reporting state F by hiding because it is likely to be in state F anyway (according to the POMDP), and by hiding it does not need to make the payment of 6 it would otherwise need to make.

To isolate the problem, suppose for a moment that payments are always possible and modify the dynamic-VCG mechanism so that step (3.) always makes payments.

Lemma 3. *When payments can be made in every period, the dynamic-VCG mechanism is interim IC and efficient with persistent and periodically-inaccessible agents.*

Proof. Property (i) in Lemma 2 holds by construction. Fix some agent i, strategy f_i, some (true) belief state bs^t, and assume agents $\neq i$ are truthful. Fix policy $\pi = \pi^*$, where π^* is optimal w.r.t. to the reported model of agent i and true model of the other agents. The payoff to agent i forward from this state is:

$$\mathbb{E}_{bs^{\geq t}}\Big[\sum_{k=t}^{K}\gamma^{k-t}r_i(bs_i^k, \pi^*(f_i(bs_i^k), bs_{-i}^k))+\sum_{k=t}^{K}\gamma^{k-t}(Q_{-i}^*(bs_{-i}^k, \pi^*(f_i(bs_i^k), bs_{-i}^k))-V_{-i}^*(bs_{-i}^k))|f_i\Big] \tag{8}$$

Here, we overload notation s.t. strategy $f_i : S_i \rightarrow S_i$ induces $f_i : BS_i \rightarrow BS_i$, with $f_i(bs_i) = f_i(s_i)$ for the corresponding state s_i if bs_i places a point mass on this state and $H(s_i)$ and $f_i(bs_i) = \phi$ otherwise. Given this, Eq. (8) is the expression for the payoff to i in bs^t, given that it follows strategy f_i, and with the expectation taken w.r.t. the distribution on future belief states given policy π. Having set this up, the rest of the proof goes through unchanged from Theorem 1. □

But agents cannot receive payments in every period, and their payoffs are not correctly aligned. Motivated by this, we consider a slight modification to dynamic-VCG:

Mechanism 3 (dynamic-VCG#). *Same as dynamic-VCG on the belief-state MDP, except that in period t in which agent i reports a message, the payment is defined as:*

$$\hat{T}_i^t(bs^t) = \sum_{k=t-\delta(t)}^{t} \frac{T_i^k(bs^k)}{\gamma^{t-k}}, \tag{9}$$

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

We now introduce a new assumption.

Assumption A2. *Each agent must eventually make any payments it owes.*

Informally, an agent can run but cannot hide for ever (or, "you must pay the piper"). Given this, the expected discounted stream of payments, given this "catch up payment," is the same as for dynamic-VCG when payments can be made in every period.

Lemma 4. *The expected payoff to agent i in dynamic-VCG#, forward from any state bs^t, for any strategy f_i is equal to that in the dynamic-VCG mechanism defined on the belief-state MDP, when payments in that mechanism can be made in every period.*

Proof. The policy is the same and the rewards received by agent i for actions in states are unchanged. Left to show is that the expected discounted stream of payments is the same. We need:

$$\mathbb{E}_{bs^{\geq t}}\Big[\sum_{k=t}^{K}\gamma^{k-t}T_i^k(f_i(bs_i^k), bs_{-i}^k)|f_i\Big] = \mathbb{E}_{bs^{\geq t}}\Big[\sum_{\substack{k=t\\ H}}^{K}\gamma^{k-t}\hat{T}_i^k(f_i(bs_i^k), bs_{-i}^k)\Big], \qquad (10)$$

where the second summation restricts to states in which agent i reports its accessibility. To show this, consider any *realization* of belief states $\mathfrak{bs}^t \dots \mathfrak{bs}^K$. We have:

$$\sum_{k=t}^{K}\gamma^{k-t}T_i^k(f_i(\mathfrak{bs}_i^k), \mathfrak{bs}_{-i}^k) = \sum_{\substack{k=t\\ H\wedge NF}}\gamma^{k-t}T_i^k(f_i(\mathfrak{bs}_i^k), \mathfrak{bs}_{-i}^k) + \sum_{\substack{k'=t\\ H\wedge F}}\gamma^{k'-t}\sum_{k=k'-\delta(k)}^{k}\frac{T_i^k(f_i(\mathfrak{bs}_i^k), \mathfrak{bs}_{-i}^k)}{\gamma^{k'-k}},$$

in which the first summation restricts to states in which agent i reports its accessibility and this is not the first time (NF) after being inaccessible (we also put the $\mathfrak{bs}^t$ state here, if accessible), and the second summation is those accessible states but where this is the first report after a being inaccessible for $\delta(k) > 0$ periods. Simple algebra completes the proof, together with assumption 2, which ensures that the final state is not inaccessible. □

Given this, we have as an immediate corollary:

Theorem 5. *Dynamic-VCG# is interim incentive compatible and efficient with persistent agents that are periodically inaccessible, and where each agent must eventually make payments owed to the center.*

By introducing the constraint that payments must always be made we avoid a manipulation in which an agent does not "re-enter" because it faces a large payment.[11]

[11] Return again to Example 3. The earlier manipulation goes away. Agent 2 in state E can no longer benefit from pretending to be inaccessible when it is in fact accessible and in state E, because it will face a payment of $-6-2$ if it makes itself accessible in period 2. But if it could avoid payments altogether, a deviation could still be useful, so Assumption A2 is key.

3.2 Dynamic Agent Population with Arrival Process

We now depart from the standard MAS model and consider a dynamically changing population, with each agent initially inaccessible, then accessible at an *arrival* period, and then becoming inaccessible again at a *departure* period forever. We conceptualize the first and last periods in which an agent is accessible as its *arrival* and *departure* periods. For motivation, we have in mind that becoming accessible corresponds to an agent learning its model, or learning of the existence of the mechanism. We assume that an agent has no reward and undergoes no state transitions while inaccessible. We continue to allow local dynamics to depend on actions after arrival. Unlike in Section 3.1, agents are not identified and cannot incur charges before arrival.

Heading for a dynamic-VCG mechanism, let us again consider the central planner's problem and formulate this as an MDP. We allow for the set of agents $N = \{1, \ldots, \infty\}$ to be unbounded. The joint MDP now defines states $s = (s_0, \{s_i\}^{i \in H(s_0)}) \in S$ where s_0 keeps sufficient history, in this case to determine both feasible actions $A(s)$ and also the dynamics for agent arrivals, and $H(s_0) \subseteq N$ is the set of accessible agents given s_0. Upon arrival, each agent is associated with a local MDP model and an initial state. This is its *arrival type.* Transitions $\tau : S \times A \to S$ are induced by an *arrival model*, $\tau_0 : S \times A \to S_0$, known to the center and defining the process by which agents become accessible, and the dynamics, $\tau_i : S_i \times A \to S_i$, for each accessible agent. The local model of an agent is augmented to include an *absorbing, inaccessible* state, so that once an agent has arrived its own model determines when it will become inaccessible. The joint reward is $r(s, a) = \sum_{i \in H(s)} r_i(s_i, a)$.

The main question is as above: *can we define an efficient mechanism in which an agent will report its state information in all periods in which it is accessible?* Consider a slight variation on the Dynamic-VCG mechanism to handle agent inaccessibility:

Mechanism 4 (Dynamic-VCG##). *At every time step t (in state s^t):*

1. *Each accessible agent i can report to the center a claim, $f_i(s_i^t)$, about its current state (including its model if this is its first report).*
2. *The center updates the joint state and selects action $a^t = \pi^*(f(s^t))$, where π^* is the optimal policy given its arrival model, so reported agent models.*
3. *The center pays each agent that sends a message a payment $T_i^t(f(s^t)) = Q_{-i}^*(f_{-i}(s_{-i}^t), \pi^*(f(s^t))) - V_{-i}^*(f_{-i}(s_{-i}^t))$, where the expected values (V_{-i}^*, Q_{-i}^*) are taken w.r.t. the reported agent models.*

The appropriate definition of *incentive compatibility* in this environment requires that agent i maximizes its payoff by truthful reporting in every *accessible* state (rather than in every state). Easier than in Section 3.1, it is the "become-accessible-once" property that makes this sufficient. Yet without an additional assumption, the dynamic-VCG## mechanism fails for a subtle reason:

Example 4. Consider an adaptation of Example 3. Suppose that agent 1 now represents an arrival type, and that there are also three other arrival types: type 2 is identical to agent 2 from Example 3, but only starting from state E forward (E is a type 2 agent's initial state), types 3 and 4 are also identical to a part of agent 2, type 3's initial state is G and type 4's initial state is H. Define an arrival process so that a single agent of type 1 always arrives in step 0 while at most one agent among types 2, 3, or 4 can arrive, and it is very likely that a type 4 agent will arrive in step 2. If an agent of type 2 arrives in step 1, then it will hide and claim to be inaccessible. The optimal policy will wait, because it likely that a type 4 agent will arrive. In period 2, the agent can truthfully report state G (posing as a type 3 agent that just arrived), and will be allocated the item for a payment of 2. This causes an efficiency loss because the item should have been allocated to agent 1 in step 1.

Lemma 1 holds unchanged in this environment. The proof of Theorem 3 also remains valid except for the very last step. The payoff to agent i in any (truly) accessible state is, as before, $V_{-i}^{\pi_{-i}}(s^t) \neq V_{-i}^{\pi_{-i}}(s_{-i}^t)$. But we now have that $\pi_{-i}^*(s^t) \neq \pi_{-i}^*(s_{-i}^t)$, and π_{-i}^* need not be independent of strategy f_i. Although we maintain Assumption A1, the probability of future agent arrivals can depend on the arrival of agent i: the model reported by agent i upon its arrival, i.e., its arrival type, or its failure to arrive, can influence the center's beliefs about subsequent arrivals. Fully general arrival dynamics introduce this new interdependence between agents. A necessary and sufficient condition for correctness of the mechanism is $V_{-i}^{\pi_{-i}}(s^t) = V_{-i}^{\pi_{-i}}(s_{-i}^t)$ for all accessible states. Here is a stronger, but appealing, condition:

Assumption A3 (CIA). *The center's arrival model, which specifies the distribution on new agent arrival types in period $t+1$ is independent of earlier arrivals.*

This immediately recovers the following theorem.

Theorem 6. *The dynamic-VCG## mechanism is interim IC and efficient in this dynamic population, become-accessible-once environment given the CIA assumption.*

The CIA assumption was implicitly made in the earlier work of Parkes and Singh (2003) (PS) in their model of "online MD." In closing we unify that earlier framework with the current framework. The *Online-VCG* mechanism of PS is payoff-equivalent to the dynamic-VCG## mechanism, when coupled with an additional assumption:

Assumption A4. *Each agent's local MDP model is deterministic.*

The effect is that the only stochasticity is due to the arrival model, so agents can report information with a single message (upon arrival). An agent's local problem is now defined by a deterministic finite-state automaton, which plays the role of type in the model of PS. We provide an interpretation of Online-VCG in an environment with discounting:

Mechanism 5 (Online-VCG-γ). *At every time step t (in state s^t):*

1. *Each accessible agent i that has yet to send a message can report to the center a claim, $f_i(s_i^t)$ about its local state and local (deterministic) model.*
2. *The center updates the joint state and selects action $a^t = \pi^*(f(s^t))$, where π^* is the optimal policy given its arrival model and reported models.*
3. *The center pays each agent that remains accessible according to its reported model a payment:*

$$\breve{T}_i^t(f(s^t)) = \begin{cases} -r_i(f_i(s_i^t), \pi^*(f(s^t))) + V^*(f(s^t)) - V^*_{-i}(f_{-i}(s^t_{-i})) & \textit{, if FT} \\ -r_i(f_i(s_i^t), \pi^*(f(s^t))) & \textit{, otherwise,} \end{cases} \tag{11}$$

where the expected values V^* and V^*_{-i} are taken w.r.t. the stochastic model of the center and given agent reports, and FT indicates that this is the period in which the agent makes its report.

The cumulative effect of the payments is that agent i pays to the center the total (reported) reward it receives for the sequence of decisions, and receives a payment of $V^*(f(s^t)) - V^*_{-i}(f_{-i}(s^t_{-i}))$ in the first period in which it announces its type. This payment is equal to the expected marginal product contributed by the agent give the stochastic model of the center and the reported types of agents.

To establish that the online-VCG-γ mechanism is IC we first present the variation on Lemma 1 that is appropriate for this environment. It is WLOG to restrict the strategy space to allow the report of just a single message, and IC simplifies to requiring that an agent's expected payoff is maximized in its true arrival period by reporting its true type immediately. This, in turn, leads to weaker conditions for IC:

Lemma 5. *A dynamic mechanism (π, T) is interim incentive compatible with dynamic, become-accessible-once agents with deterministic local MDPs given the CIA assumption if:*

i) policy π is optimal given reported models, and

ii) agent i's expected payoff (w.r.t. the true model) in the state s^t in which it first becomes accessible is $V^\pi(s^t|f_i) - C_i(s^t)$, where $C(s^t)$ is a constant, and independent of strategy f_i.

(Proof omitted because it follows the same pattern as for the earlier variants on

Lemma 1.)

Theorem 7. *Online-VCG-γ is interim IC and efficient in this dynamic population, become-accessible-once setting given the CIA assumption and agents with deterministic local MDPs.*

Proof. Property (i) in Lemma 5 holds by construction. Fix some agent i, strategy f_i, state s^t in which agent i arrives, and assume agents $\neq i$ are truthful. Fix policy $\pi = \pi^*$, where π^* is optimal. The payoff to agent i forward from this state is:

$$\mathbb{E}_{s^{\geq t}}\Big[\sum_{k=t}^{K}\gamma^{k-t} r_i(s_i^k, \pi^*(f_i(s_i^k), s_{-i}^k)) + \sum_{\substack{k=t\\ H\wedge FT}}^{K} \gamma^{k-t}\left(V^*(f_i(s_i^k), s_{-i}^k) - V_{-i}^*(s_{-i}^k)\right) - \sum_{\substack{k=t\\ H\wedge\neg FT}}^{K} \gamma^{k-t} r_i(f_i(s_i^k), \pi^*(f_i(s_i^k), s_{-i}^k))|f_i\Big],$$

where H indicates the agent reports that it is accessible, and FT indicates the period is the one in which agent i reports its type. Label the four terms $\{\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}\}$ and introduce the following two terms:

$$\mathbb{E}_{s\geq t}\Big[\sum_{\substack{k=t\\ BA}}^{K}\gamma^{k-t} r_{-i}(s_{-i}^k) - \sum_{\substack{k=t\\ BA}}^{K}\gamma^{k-t} r_{-i}(s_{-i}^k)\Big], \tag{12}$$

labeled $\mathbf{E}$ and $\mathbf{F}$ respectively, and with BA ("before arrival") indicating that these terms are defined on states s^k for which agent i has not reported its accessibility. We complete the proof by concluding that the payoff to agent i equals

$$V^{\pi}(s^t|f_i) - V_{-i}^*(s_{-i}^t), \tag{13}$$

as required with the first term coming from $\mathbf{A} + \mathbf{E} + \mathbf{B} - \mathbf{D}$ and the second term coming from $\mathbf{F} + \mathbf{C}$. □

One reason to adopt Online-VCG-γ rather than dynamic-VCG## in this special environment is that the payments require solving $V_{-i}^*(f_i(s_{-i}^t))$ only once for each agent arrival, whereas in dynamic-VCG## this problem must be solved in every period in which the agent remains accessible according to its report.

4 Potential Applications

The dynamic-VCG mechanism is applied by Bergemann and Välimäki (2006) to a multi-agent variant on the multi-armed bandit problem (see also Cavallo et al. (2006)). In that environment it provides optimal, coordinated planning with local Markov-chain models, including, for instance, optimal Bayesian learning. The dynamic-VCG# variation also applies when agents receive "interrupts" and are periodically inaccessible. The dynamic-VCG## variation extends to multi-agent systems with dynamic populations, for instance when agents with stochastic local state compete for shared resources, and encompasses the online MD environment of Parkes and Singh (2003), in which each agent has an arrival and departure and learns its “type” upon arrival.

References

[Athey and Segal, 2007] Susan Athey and Ilya Segal. An efficient dynamic mechanism. Working paper, http://www.stanford.edu/ isegal/agv.pdf, 2007.

[Bergemann and Valimaki, 2006] Dirk Bergemann and Juuso Valimaki. Efficient dynamic auctions. Cowles Foundation Discussion Paper 1584, http://cowles.econ.yale.edu/P/cd/d15b/d1584.pdf, 2006.

[Cavallo *et al.*, 2006] Ruggiero Cavallo, David C. Parkes, and Satinder Singh. Optimal coordinated planning amongst self-interested agents with private state. In *Proceedings of the Twenty-second Annual Conference on Uncertainty in Artificial Intelligence (UAI'06)*, 2006.

[Cremer and McLean, 1985] J. Cremer and R. McLean. Optimal selling strategies under uncertainty for a discriminating monopolist when demands are interdependent. *Econometrica*, 53:345–361, 1985.

[Jackson, 2003] Matthew O. Jackson. Mechanism theory. In Ulrich Derigs, editor, *The Encyclopedia of Life Support Systems*. EOLSS Publishers, 2003.

[Jehiel and Moldovanu, 2001] Philippe Jehiel and Benny Moldovanu. Efficient design with interdependent valuations. *Econometrica*, 69:1237–1259, 2001.

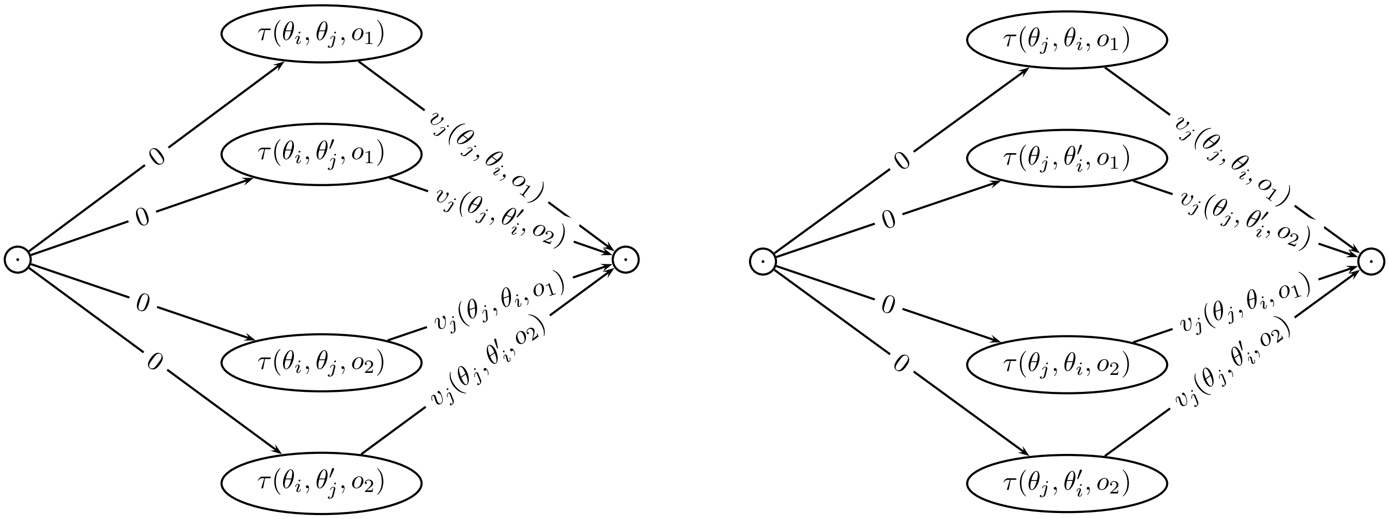
[Kaelbling *et al.*, 1996] Leslie Pack Kaelbling, Michael L. Littman, and Andrew W. Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

[Lavi and Nisan, 2000] Ron Lavi and Noam Nisan. Competitive analysis of incentive compatible on-line auctions. In *Proc. 2nd ACM Conf. on Electronic Commerce (EC-00)*, pages 233–241, 2000.

[Parkes and Singh, 2003] David C. Parkes and Satinder Singh. An MDP-based approach to Online Mechanism Design. In *Proc. 17th Annual Conf. on Neural Information Processing Systems (NIPS'03)*, 2003.

[Parkes, 2007] David C Parkes. On-line mechanisms. In Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay Vazirani, editors, *Algorithmic Game Theory*, chapter 16. Cambridge University Press, 2007.

5 Appendix



$\tau(\theta_i, \theta_j, o_1)$
$\tau(\theta_i, \theta'_j, o_1)$
$\tau(\theta_i, \theta_j, o_2)$
$\tau(\theta_i, \theta'_j, o_2)$
0
0
0
0
$v_j(\theta_j, \theta_i, o_1)$
$v_j(\theta_j, \theta'_i, o_2)$
$v_j(\theta_j, \theta_i, o_1)$
$v_j(\theta_j, \theta'_i, o_2)$
$\tau(\theta_j, \theta_i, o_1)$
$\tau(\theta_j, \theta'_i, o_1)$
$\tau(\theta_j, \theta_i, o_2)$
$\tau(\theta_j, \theta'_i, o_2)$
0
0
0
0
$v_j(\theta_j, \theta_i, o_1)$
$v_j(\theta_j, \theta'_i, o_2)$
$v_j(\theta_j, \theta_i, o_1)$
$v_j(\theta_j, \theta'_i, o_2)$

(a) Agent i's MDP.

(b) Agent j's MDP.

Figure 2: Mapping of a static interdependent values problem to a dynamic problem with transitions that depend on other agents' states and rewards that don't.