Influence-Optimistic Local Values
for Multiagent Planning — Extended Version

Frans A. Oliehoek
University of Amsterdam
University of Liverpool
fao@liverpool.ac.uk

Matthijs T. J. Spaan
Delft University of Technology
The Netherlands
m.t.j.spaan@tudelft.nl

Stefan Witwicki
Swiss Federal Institute
of Technology (EPFL)
stefan.witwicki@epfl.ch

ABSTRACT

Recent years have seen the development of methods for multiagent planning under uncertainty that scale to tens or even hundreds of agents. However, most of these methods either make restrictive assumptions on the problem domain, or provide approximate solutions *without* any guarantees on quality. Methods in the former category typically build on heuristic search using upper bounds on the value function. Unfortunately, no techniques exist to compute such upper bounds for problems with non-factored value functions. To allow for meaningful benchmarking through measurable quality guarantees on a very general class of problems, this paper introduces a family of *influence-optimistic* upper bounds for factored decentralized partially observable Markov decision processes (Dec-POMDPs) that do not have factored value functions. Intuitively, we derive bounds on very large multiagent planning problems by subdividing them in sub-problems, and at each of these sub-problems making optimistic assumptions with respect to the influence that will be exerted by the rest of the system. We numerically compare the different upper bounds and demonstrate how we can achieve a non-trivial guarantee that a heuristic solution for problems with hundreds of agents is close to optimal. Furthermore, we provide evidence that the upper bounds may improve the effectiveness of heuristic influence search, and discuss further potential applications to multiagent planning.

1. INTRODUCTION

Planning for multiagent systems (MASs) under uncertainty is an important research problem in artificial intelligence. The decentralized partially observable Markov decision process (Dec-POMDP) is a general principled framework for addressing such problems. Many recent approaches to solving Dec-POMDPs propose to exploit *locality of interaction* [22] also referred to as *value factorization* [16]. However, without making very strong assumptions, such as transition and observation independence [3], there is no strict locality: in general the actions of any agent may affect the rewards received in a different part of the system, even if that agent and the origin of that reward are (spatially) far apart. For instance, in a traffic network the actions taken in one part of the network will eventually influence the rest of the network [26].

A number of approaches have been proposed to generate solutions for large MASs [40, 48, 27, 47, 9, 36]. However, these heuristic methods come without guarantees. In fact, since it has been shown that approximation (given some ϵ, finding a solution with value within ϵ of optimal) of Dec-POMDPs is NEXP-complete [31], it is unrealistic to expect to find general, scalable methods that have such guarantees. However, the lack of guarantees also makes it difficult to meaningfully interpret the results produced by heuristic methods. In this work, we mitigate this issue by proposing a novel set of techniques that can be used to provide upper bounds on the performance of large *factored* Dec-POMDPs.

More generally, the ability to compute upper bounds is important for numerous reasons: 1) As stated above, they are crucial for a meaningful interpretation of the quality of heuristic methods. 2) Such knowledge of performance gaps is crucial for researchers to direct their focus to promising areas. 3) Such knowledge is also crucial for understanding which problems seem simpler to approximate than others, which in turn may lead to improved theoretical understanding of different problems. 4) Knowledge about the performance gap of the leading heuristic methods can also accelerate their real-world deployment, e.g., when their performance gap is proven to be small over sampled domain instances, or when the selection of which heuristic method to deploy is facilitated by clarifying the trade-off of computation and closeness to optimality. 5) Upper bounds on achievable value without communication may guide decisions on investments in communication infrastructure. 6) Last, but not least, these upper bounds can directly be used in current and future heuristic search methods, as we will discuss in some more detail at the end of this paper.

Computing upper bounds on the achievable value of a planning problem typically involves relaxing the original problem by making some optimistic assumptions. For instance, in the case of Dec-POMDPs typical assumptions are that the agents can communicate or observe the true state of the system [11, 35, 32, 25]. By exploiting the fact that transition and observation dependence leads to a value function that is *additively factored* into a number of small components (we say that the value function is 'factored', or that the setting exhibits 'value

factorization'), such techniques have been extended to compute upper bounds for so-called *network-distributed POMDPs (ND-POMDPs)* with many agents. This has greatly increased the size of the problems that can be solved [38, 19, 9]. Unfortunately, assuming both transition and observation independence (or, more generally, value factorization) narrows down the applicability of the model, and no techniques for computing upper bounds for more general factored Dec-POMDPs with many agents are currently known.

We address this problem by proposing a general technique for computing what we call *influence-optimistic* upper bounds. These are upper bounds on the achievable value in large-scale MASs formed by computing *local* influence-optimistic upper bounds on the value of sub-problems that consist of small subsets of agents and state factors. The key idea is that if we make optimistic assumptions about how the rest of the system will influence a sub-problem, we can decouple it and effectively compute a local upper bound on the achievable value. Finally, we show how these local bounds can be combined into a *global* upper bound. In this way, the major contribution of this paper is that it shows how we can compute *factored upper bounds* for models that *do not admit factored value functions.*

We empirically evaluate the utility of influence-optimistic upper bounds by investigating the quality guarantees they provide for heuristic methods, and by examining their application in a heuristic search method. The results show that the proposed bounds are tight enough to give meaningful quality guarantees for the heuristic solutions for factored Dec-POMDPs with hundreds of agents.[1] This is a major accomplishment since previous approaches that provide guarantees 1) have required very particular structure such as transition and observation independence [3, 2, 38, 9] or 'transition-decoupledness' combined with very specific interaction structures (transitions of an agent can be affected in a *directed* fashion and only by a small subset of other agents) [42], and 2) have not scaled beyond 50 agents. In contrast, this paper demonstrates quality bounds in settings of hundreds of agents that all influence each other via their actions.

This paper is organized as follows. First, Section 2 describes the required background by introducing the factored Dec-POMDP model. Next, Section 3 describes the sub-problems that form the basis of our decomposition scheme. Section 4 proposes local influence-optimistic upper bounds for such sub-problems together with the techniques to compute them. Subsequently, Section 5 discusses how these local upper bounds can be combined into a global upper bound for large problems with many agents. Section 6 empirically investigates the merits of the proposed bounds. Section 7 places our work in the context of related work in more detail, and Section 8 concludes.

---

[1] In the paper, we use the word 'tight' for its (empirical) meaning of "close to optimal", not for its (theoretical CS) meaning of "coinciding with the best possible bound".
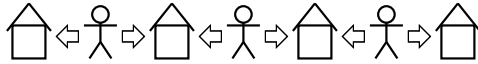

Figure 1: The FireFightingGraph problem.

2. BACKGROUND

In this paper we focus on factored Dec-POMDPs [26], which are Dec-POMDPs where the transition and observation models can be represented compactly as a *two-stage dynamic Bayesian network* (2DBN) [4]:

Definition 1. A *factored Dec-POMDP* is a tuple $\mathcal{M} = \langle \mathcal{D}, \mathcal{A}, \mathcal{O}, \mathcal{X}, T, O, \mathcal{R}, b^0 \rangle$, where:

- $\mathcal{D} = \{1, \ldots, |\mathcal{D}|\}$ is the set of agents.
- $\mathcal{A} = \bigotimes_{i \in \mathcal{D}} \mathcal{A}_i$ is the set of joint actions a.
- $\mathcal{O} = \bigotimes_{i \in \mathcal{D}} \mathcal{O}_i$ is the set of joint observations o.
- $\mathcal{X} = \{X^1, \ldots, X^m\}$ is a set of state variables, or *factors*, that take values $Dom(X^k)$ and thus span the set of states $\mathcal{S} = \bigotimes_{X^k \in \mathcal{X}} Dom(X^k)$.
- $T(s'|s,a)$ is the transition model which is specified by a set of conditional probability tables (CPTs), one for each factor.
- $O(o|a,s')$ is the observation model, specified by a CPT per agent.
- $\mathcal{R} = \{R^1, \ldots, R^\rho\}$ is a set of *local* reward functions.
- b^0 is the (factored) initial state distribution.

Each local reward function R^l has a *state factor scope* $\mathcal{X}(l) \subseteq \mathcal{X}$ and *agent scope* $\mathcal{D}(l) \subseteq \mathcal{D}$ over which is it is defined: $R^l(x_{\mathcal{X}(l)}, a_{\mathcal{D}(l)}, x'_{\mathcal{X}(l)}) \in \mathbb{R}$. These local reward functions form the global immediate reward function via addition. We slightly abuse notation and overload l to denote both an index into the set of reward functions, as well as the corresponding scopes:

$$R(s,a,s') \triangleq \sum_{l \in \mathcal{R}} R^l(x_l, a_l, x'_l).$$

Every Dec-POMDP can be converted to a factored Dec-POMDP, but the additional structure that a factored model specifies is most useful when the problem is *weakly coupled*, meaning that there is sufficient conditional independence in the 2DBN and that the scopes of the reward functions are small.

For instance, Fig. 1 shows the FIREFIGHTINGGRAPH (FFG) problem [27], which we adopt as a running example. This problem defines a set of $|\mathcal{D}| + 1$ houses, each with a particular 'fire level' indicating if the house is burning and with what intensity. Each agent can fight fire at the house to its left or right, making observations of flames (or no flames) at the visited house. Each house has a local reward function associated with it, which depends on the next-stage fire-level,[2] as illustrated in Fig. 2(left) which shows the 2DBN for a 4-agent instantiation of FFG. The figure shows that the connections

---

[2] FFG has rewards of form $R^l(x'_l)$, but we support $R^l(x_l, a_l, x'_l)$ in general.
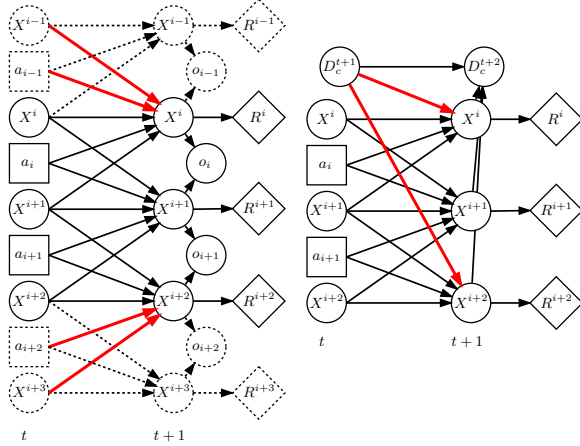
Figure 2: Left: A 2-agent sub-problem within 4-agent FFG. Right: the corresponding IASP.

are *local* but there is no *transition independence* [3] or *value factorization* [16, 42]: all houses and agents are connected such that, over time, actions of each agent can influence the entire system. While FFG is a stylized example, such locally-connected systems can be found in applications as traffic control [47] or communication networks [29, 12, 18].

This paper focuses on problems with a finite horizon h such that $t = 0, \ldots, h - 1$. A policy π_i for an agent i specifies an action for each observation history $\vec{o}_i^t = (o_i^1, \ldots, o_i^t)$. The task of planning for a factored Dec-POMDP entails finding a joint policy $\pi = \langle \pi_1, \ldots, \pi_{|\mathcal{D}|} \rangle$ with maximum *value*, i.e., expected sum of rewards:

$$V(\pi) \triangleq \mathbf{E} \left[\sum_{t=0}^{h-1} R(s,a,s') \mid b^0, \pi \right].$$

Such an optimal joint policy is denoted π^*.[3]

In recent years, a number of methods have been proposed to find approximate solutions for factored Dec-POMDPs with many agents [30, 16, 40, 27, 47] but none of these methods are able to give guarantees with respect to the solution quality (i.e., they are *heuristic* methods), leaving the user unable to confidently gauge how well these methods perform on their problems. This is a principled problem; even finding an ϵ-approximate solution is NEXP-complete [31], which implies that general and efficient approximation schemes are unlikely to be found. In this paper, we propose a way forward by trying to find instance-specific upper bounds in order to provide information about the solution quality offered by heuristic methods.

3. SUB-PROBLEMS AND INFLUENCES

The overall approach that we take is to divide the problem into sub-problems (defined here), compute over-

---

[3] We omit the '*' on values; all values are assumed to be optimal with respect to their given arguments.

estimations of the achievable value for each of these sub-problems (discussed in Section 4) and combine those into a global upper bound (Section 5).

3.1 Sub-Problems (SPs)

The notion of a sub-problem generalizes the concept of a *local-form model (LFM)* [28] to multiple agents and reward components. We give a relatively concise description of this formalization, for more details, please see [28].

Definition 2. A *sub-problem (SP)* $\mathcal{M}_c$ of a factored Dec-POMDP $\mathcal{M}$ is a tuple $\mathcal{M}_c = \langle \mathcal{M}, \mathcal{D}', \mathcal{X}', \mathcal{R}' \rangle$, where $\mathcal{D}' \subset \mathcal{D}, \mathcal{X}' \subset \mathcal{X}, \mathcal{R}' \subset \mathcal{R}$ denote subsets of agents, state factors and local reward functions.

An SP inherits many features from $\mathcal{M}$: we can define local states $x_c \in \bigotimes_{X \in \mathcal{X}'}$ and the subsets $\mathcal{D}', \mathcal{X}', \mathcal{R}'$ induce local joint actions $\mathcal{A}_c = \bigotimes_{i \in \mathcal{D}'} \mathcal{A}_i$, observations $\mathcal{O}_c = \bigotimes_{i \in \mathcal{D}'} \mathcal{O}_i$, and rewards

$$R_c(x_c, a_c, x_c') \triangleq \sum_{l \in \mathcal{R}'} R^l(x_l, a_l, x_l'). \tag{1}$$

However, this is generally not enough to end up with a fully specified, but smaller, factored Dec-POMDP. This is illustrated in Fig. 2(left), which shows the 2DBN for a sub-problem of FFG involving two agents and three houses (dependence of observations o_i on actions a_i is not displayed). The figure shows that state factors $X \in \mathcal{X}'$ (in this case X^i and X^{i+2}) can be the target of arrows pointing into the sub-problem from the non-modeled (dashed) part. We refer to such state factors as *non-locally affected factors (NLAFs)* and denote them xn_c^k, where c indexes the SP and k indexes the factor. The other state factors in $\mathcal{X}'$ are referred to as *only-locally affected factors (OLAFs)* xl_c^k. The figure clearly shows that the transition probabilities are not well-defined since the NLAFs depend on the sources of the highlighted *influence links*. We refer to these sources as *influence sources* $u_c^{t+1} = \langle y_u^t, a_u^t \rangle$ (in this case $y_u^t = \langle X^{i-1}, X^{i+3} \rangle$ and $a_u^t = \langle a_{i-1}^t, a_{i+2}^t \rangle$). This means that an SP c has an *underspecified transition model*: $T_c(x_c^{t+1} | x_c^t, a_c^t, u_c^{t+1})$.

3.2 Structural Assumptions

In the most general form, the observation and reward model could also be underspecified. In order to simplify the exposition, we make two assumptions on the structure of an SP:

1. For all included agents $i \in \mathcal{D}'$, the state factors that can influence its observations (i.e., ancestors of o_i in the 2DBN) are included in $\mathcal{M}_c$.
2. For all included reward components $R^l \in \mathcal{R}'$, the state factors and actions that influence R^l are included in $\mathcal{M}_c$.

That is, we assume that SPs exhibit generalized forms of *observation independence*,

$$O_c(o_c | a_c, x_c') \triangleq \Pr(o_c | a_c, x_c') = \Pr(o_c | a, s'),$$

and *reward independence* (cf. (1)). These are more general notions of observation and reward independence than used in previous work on TOI-Dec-MDPs [3] and ND-POMDPs [22], since we allow overlap on state factors that can be influenced by the agents themselves.[4]

Crucially, however, we do not assume any form of transition independence (for instance, the sets $\mathcal{X}'$ of SPs can overlap), nor do we assume any of the transition-decoupling (i.e., TD-POMDP [43]) restrictions. That is, we *neither* restrict which node types can affect 'private' nodes; nor do we disallow concurrent interaction effects on 'mutually modeled' nodes.

This means that assumptions 1 and 2 (above) that we do make are without loss of generality: it is possible to make any Dec-POMDP problem satisfy them by introducing additional (dummy) state factors.[5]

3.3 Influence-Augmented SPs

An LFM can be transformed into a so-called *influence-augmented local model*, which captures the influence of the policies and parts of the environment that are not modeled in the local model [28]. Here we extend this approach to SPs, thus leading to *influence-augmented sub-problems (IASPs)*.

Intuitively, the construction of an IASP consists of two steps: 1) capturing the influence of the non-modeled parts of the problem (given $\pi_{\neq c}$ the policies of non-modeled agents) in an *incoming influence point* $I_{\to c}(\pi_{-c})$, and 2) using this $I_{\to c}$ to create a model with a transformed transition model $T_{I_{\to c}}$ and no further dependence on the external problem.

Step (1) can be done as follows: an *incoming influence point* can be specified as an *incoming influence* $I_{\to c}^t$ for each stage: $I_{\to c} = \left(I_{\to c}^1, \ldots, I_{\to c}^h \right)$. Each such $I_{\to c}^{t+1}$ corresponds to the influence that the SP experiences at stage $t+1$, and thus specifies the conditional probability distribution of the influence sources $u_c^{t+1} = \left\langle y_u^t, a_u^t \right\rangle$. That is, assuming that the influencing agents use deterministic policies $\pi_u = \langle \pi_i \rangle_{i \in u}$ that map observation histories to actions, $I_{\to c}^{t+1}$ is the conditional probability distribution given by

$$I(u_c^{t+1} \mid D_c^{t+1}) = \sum_{\vec{o}_u^t} \mathbf{1}_{\{a_u^t = \pi_u(\vec{o}_u^t)\}} \Pr(y_u^t, \vec{o}_u^t | D_c^{t+1}, b^0, \pi_{\neq c}),$$

where $\mathbf{1}_{\{\cdot\}}$ is the Kronecker Delta function, and D_c^{t+1} the *d-separating set* for $I_{\to c}^{t+1}$: the history of a subset of all the modeled variables that d-separates the modeled variables from the non-modeled ones.[6]

Step (2) involves replacing the CPTs for all the NLAFs by the CPTs *induced* by $I_{\to c}$.

Definition 3. Let $xn_c^{k,t+1}$ be an NLAF (with index k), and u_c^{t+1} (the instantiation of) the corresponding influence sources. Given the influence $I_{\to c}^{t+1}(\pi_{\neq c})$, and its d-separating set D_i^{t+1}, we define the *induced CPT* for $xn_c^{k,t+1}$ as the CPT that specifies probabilities:

$$p_{I_{\to c}^{t+1}}(xn_c^{k,t+1} | x_c^t, D_c^{t+1}, a_c^t) = \sum_{u_c^{t+1} = \langle y_u^t, a_u^t \rangle} \Pr(xn_c^{k,t+1} | x_c^t, a_c^t, u_c^{t+1}) I(u_c^{t+1} | D_c^{t+1}). \quad (2)$$

Finally, we can define the IASP.

Definition 4. An *influence-augmented SP (IASP)* $\mathcal{M}_c^{IA} = \langle \mathcal{M}_c, I_{\to c} \rangle$ for an SP $\mathcal{M}_c = \langle \mathcal{M}, \mathcal{D}', \mathcal{X}', \mathcal{R}' \rangle$ is a factored Dec-POMDP with the following components:

- The agents (implying the actions and observations) from the respective subproblem participate: $\bar{\mathcal{D}} = \mathcal{D}'$.
- The set of state factors is $\bar{\mathcal{X}} = \mathcal{X}' \cup \{D_c\}$ such that states $\bar{x}_c^t = \langle x_c^t, D_c^{t+1} \rangle$ specify a local state of the SP, as well as the d-separating set D_i^{t+1} for the next-stage influences.
- Transitions are specified as follows: For all OLAFs x_c^k we take the CPTs from the factored Dec-POMDP $\mathcal{M}$, but for all NLAFs we take the induced CPTs, leading to an influence-augmented transition model which is the product of CPTs of OLAFs and NLAFS:

$$\bar{T}_{I_{\to c}}(x_c^{t+1} | \langle x_c^t, D_c^{t+1} \rangle, a_c^t) = \Pr(xl_c^{t+1} | x_c^t, a_c^t) \sum_{u_c^{t+1} = \langle y_u^t, a_u^t \rangle} \Pr(xn_c^{t+1} | x_c^t, a_c^t, u_c^{t+1}) I(u_c^{t+1} | D_c^{t+1}). \quad (3)$$

(Note that x_c^t, a_c^t, x_c^{t+1} and D_c^{t+1} together uniquely specify D_c^{t+2}).

- The observation model $\bar{O}$ follows directly from O (from $\mathcal{M}$).
- The reward is identical to that of the SP: $\bar{\mathcal{R}} = \mathcal{R}'$.

Fig. 2(right) illustrates the IASP for FFG. It shows how the d-separating set acts as a parent for all NLAFs, thus replacing the dependence on the external part of the problem.

We write $V_c(\pi)$ for the value that would be realized for the reward components modeled in sub-problem c, under a given joint policy π:

$$V_c(\pi) \triangleq \mathbf{E} \left[\sum_{t=0}^{h-1} R_c^t(s, a, s') \mid b^0, \pi \right].$$

As one can derive, given the policies of other agents $\pi_{\neq c}$, $V_c(I_{\to c}(\pi_{\neq c}))$, the value of the optimal solution of an IASP constructed for the influence corresponding to $\pi_{\neq c}$, is equal to the best-response value:

$$V_c^{BR}(\pi_{\neq c}) \triangleq \max_{\pi_c} V_c(\pi_c, \pi_{\neq c}) = V_c(I_{\to c}(\pi_{\neq c})). \quad (4)$$

This extends the result in [28] to multiagent SPs.

4. LOCAL UPPER BOUNDS

In this section we present our main technical contribution: the machinery to compute *influence-optimistic upper bounds (IO-UBs)* for the value of sub-problems. In order to properly define this class of upper bound, we first define the *locally-optimal value*:

Definition 5. The *locally-optimal value* for an SP c,

$$V_c^{LO} \triangleq \max_{\pi_{\neq c}} V_c^{BR}(\pi_{\neq c}) = \max_{\pi_{\neq c}} V_c(I_{\to c}(\pi_{\neq c})), \quad (5)$$

is the local value (considering only the rewards R_c) that can be achieved when all agents use a policy selected to optimize this local value. We will denote the maximizing argument by $\pi_{\neq c}^{LO}$.

Note that $V_c^{LO} \geq V_c(\pi^*)$—the value for the rewards R_c under the optimal joint policy π^*—since π^* optimizes the sum of all local reward functions: it might be optimal to sacrifice some reward R_c if it is made up by higher rewards outside of the sub-problem.

V_c^{LO} expresses the maximal value achievable under a *feasible* incoming influence point; i.e., it is optimistic about the influence, but maintains that the influence is feasible. Computing this value can be difficult, since computing influences and subsequently constructing and optimally solving an IASP can be very expensive in general. However, it turns out computing upper bounds to V_c^{LO} can be done more efficiently, as discussed in Section 4.4.

The IO-UBs that we propose in the remainder of this section upper-bound V_c^{LO} by relaxing the requirement of the incoming influence being feasible, thus allowing for more efficient computation. We present three approaches that each overestimate the value by being optimistic with respect to the assumed influence, but that differ in the additional assumptions that they make.

4.1 A Q-MMDP Approach

The first approach we consider is called *influence-optimistic Q-MMDP (IO-Q-MMDP)*. Like all the heuristics we introduce, it assumes that the considered SP will receive the most optimistic (possibly infeasible) influence. In addition, it assumes that the SP is fully observable such that it reduces to a local multiagent MDP (MMDP) [5]. In other words, this approach resembles Q-MMDP [35, 25], but is applied to an SP, and performs an influence-optimistic estimation of value.[7] IO-Q-MMDP makes, in addition to influence optimism, another overestimation due to its assumption of full observability. While this negatively affects the tightness of the upper bound, it has as the advantage that its computational complexity is relatively low.

Formally, we can describe IO-Q-MMDP as follows. In the first phase, we apply dynamic programming to

[7]What we have termed "Q-MMDP" has been referred to in past work as "Q-MDP"; we add the extra M emphasize the presence of multiple agents.

compute the action-values for all local states:

$$Q(x_c^t, a_c^t) = \max_{u_c^{t+1}} \sum_{x_c^{t+1}} \Pr(xl_c^{t+1}|x_c^t, a_c^t) \Pr(xn_c^{t+1}|x_c^t, a_c^t, u_c^{t+1})$$
$$\left[R_c(x_c^t, a_c^t, x_c^{t+1}) + \max_{a_c^{t+1}} Q(x_c^{t+1}, a_c^{t+1}) \right]. \quad (6)$$

Comparing this equation to (3), it is clear that this equation is optimistic with respect to the influence: it selects the sources u_c^{t+1} in order to select the most beneficial transition probabilities. In the second phase, we use these values to compute an upper bound:

$$\hat{V}_c^M \triangleq \max_{a_c} \sum_{x_c} b^0(x_c) Q(x_c, a_c).$$

This procedure is guaranteed to yield an upper bound to the locally-optimal value for the SP.

Theorem 1. *IO-Q-MMDP yields an upper bound to the locally-optimal value:* $V_c^{LO} \leq \hat{V}_c^M$.

Proof. An inductive argument easily establishes that, due to the maximization it performs, (6) is at least as great as the Q-MMDP value (for all D_c^{t+1}) of *any* feasible influence, given by:

$$Q_c^{MMDP}(\langle x_c^t, D_c^{t+1} \rangle, a_c^t) = \sum_{x_c^{t+1}} \bar{T}_{I \to c}(x_c^{t+1}|\langle x_c^t, D_c^{t+1} \rangle, a_c^t)$$
$$\left[R_c(x_c^t, a_c^t, x_c^{t+1}) + \max_{a_c^{t+1}} Q(\langle x_c^{t+1}, D_c^{t+2} \rangle, a_c^{t+1}) \right]. \quad (7)$$

Therefore the value computed in (6) is at least as great as the Q-MMDP value (7) induced by $\pi_{\neq c}^{LO}$ (the maximizing argument of (5)), for all x_c^t, D_c^{t+1}, a_c^t. This directly implies

$$\hat{V}_c^M \geq V_c^{MMDP}(I_{\to c}(\pi_{\neq c}^{LO})),$$

Moreover, it is well known that, for any Dec-POMDP, the Q-MMDP value is an upper bound to its value [35], such that

$$V_c^{MMDP}(I_{\to c}(\pi_{\neq c}^{LO})) \geq V_c(I_{\to c}(\pi_{\neq c}^{LO})).$$

We can conclude that $\hat{V}_c^M$ is an upper bound to the Dec-POMDP value of the IASP induced by $\pi_{\neq c}^{LO}$:

$$\hat{V}_c^M \geq V_c(I_{\to c}(\pi_{\neq c}^{LO})) = \max_{\pi_{\neq c}} V_c(I_{\to c}(\pi_{\neq c})) = V_c^{LO},$$

with the identities given by (5), thus proving the theorem. □

The upshot of (6) is that there are no dependencies on d-separating sets and incoming influences anymore: the IO assumption effectively eliminates these dependencies. As a result, there is no need to actually construct the IASPs (that potentially have a very large-state-space) if all we are interested in is an upper bound.

4.2 A Q-MPOMDP Approach

The IO-Q-MMDP approach of the previous section introduces overestimations through influence-optimism as well as by assuming full observability. Here we tighten the upper bound by weakening the second assumption. In particular, we propose an upper bound based on the underlying *multiagent POMDP (MPOMDP)*.

An MPOMDP [21, 1] is partially observable, but assumes that the agents can freely communicate their observations, such that the problem reduces to a special type of centralized model in which the decision maker (representing the entire team of agents) takes joint actions, and receives joint observations. As a result, the optimal value for an MPOMDP is analogous to that of a POMDP:

$$Q(b^t, a^t) = R(b^t, a^t) + \sum_{o^{t+1}} \Pr(o^{t+1}|b^t, a^t)V(b^{t+1}), \quad (8)$$

where b^{t+1} is the joint belief resulting from performing Bayesian updating of b^t given a^t and o^{t+1}.

Using the value function of the MPOMDP solution as a heuristic (i.e., an upper bound) for the value function of a Dec-POMDP is a technique referred to as Q-MPOMDP [32, 25]. Here we combine this approach with optimistic assumptions on the influences, leading to *influence-optimistic Q-MPOMDP (IO-Q-MPOMDP)*.

In case that the influence on an SP is fully specified, (8) can be readily applied to the IASP. However, we want to deal with the case where this influence is not specified. The basic, conceptually simple, idea is to move from the *influence-optimistic* MMDP-based upper bounding scheme from in Section 4.1 to one based on MPOMDPs. However, it presents a technical difficulty, since it is not directly obvious how to extend (6) to deal with partial observability. In particular, in the MPOMDP case as given by (8), the state x_c^t is replaced by a belief over such local states and the influence sources u_c^{t+1} affect the value by both manipulating the transition and observation probabilities, as well as the resulting beliefs.

To overcome these difficulties, we propose a formulation that is not directly based on (8), but that too makes use of 'back-projected value vectors'. That is, it is possible to rewrite the optimal MPOMDP value function as:[8]

$$Q(b^t, a^t) = b^t \cdot r_a + \gamma \sum_{o^{t+1}} \max_{\nu^{ao} \in \mathcal{V}^{ao}} b^t \cdot \nu^{ao}, \quad (9)$$

where $\cdot$ denotes inner product and where $\nu^{ao} \in \mathcal{V}^{ao}$ are the back-projections of value vectors $\nu \in \mathcal{V}^{t+1}$:

$$\nu^{ao}(s^t) \triangleq \sum_{s^{t+1}} O(o^{t+1}|a^t, s^{t+1})T(s^{t+1}|s^t a^t)\nu(s^{t+1}). \quad (10)$$

(Please see, e.g., [34, 33] for more details.)

The key insight that enables carrying influence-optimism to the MPOMDP case is that this back-projected

form (10) does allow us to take the maximum with respect to unspecified influences. That is, we define the *influence-optimistic back-projection* as:

$$\nu_{IO}^{ao}(x_c^t) \triangleq \max_{u_c^{t+1}} \sum_{x_c^{t+1}} O(o_c^{t+1}|a_c, x_c^{t+1})$$

$$\Pr(xn_c^{t+1}|x_c^t, a_c, u_c^{t+1})\Pr(xl_c^{t+1}|x_c^t, a_c)\nu_{IO}(x_c^{t+1}). \quad (11)$$

Since this equation does not depend in any way on the d-separating sets and influence, we can completely avoid generating large IASPs. As for implementation, many POMDP solution methods [6, 13] are based on such back-projections and therefore can be easily modified; all that is required is to substitute these the back-projections by their modified form (11). When combined with an *exact* POMDP solver, such influence-optimistic back-ups will lead to an upper bound $\hat{V}_c^P$, to which we refer as *IO-Q-MPOMDP*, on the locally-optimal value.

To formally prove this claim, we will need to discriminate a few different types of value, and associated constructs. Let us define:

- $\bar{b}_{I \to c}^t(\langle x_c^t, D_c^{t+1} \rangle)$, an MPOMDP belief for the IASP induced by an arbitrary influence $I_{\to c}$,

- $V_{I \to c}^P$, the optimal value function when the IASP is solved as an MPOMDP, such that $V_{I \to c}^P(\bar{b}_{I \to c}^t)$ is the value of $\bar{b}_{I \to c}^t$ and $V_c^{MPOMDP}(I_{\to c}) \triangleq V_{I \to c}^P(\bar{b}_{I \to c}^0)$. $V_{I \to c}^P$ is represented using vectors $\nu \in \mathcal{V}$.

- $\bar{b}_{IO}^t(x_c^t)$, an arbitrary distribution over x_c^t that can be thought of as the MPOMDP belief for the 'influence optimistic SP'.[9]

- $V_{IO,c}^P$, the value function computed by an (exact) influence-optimistic MPOMDP method, that assigns a value $V_{IO,c}^P(\bar{b}_{IO}^t)$ to any $\bar{b}_{IO}^t$. $V_{IO,c}^P$ is represented using vectors $\nu_{IO} \in \mathcal{V}_{IO}$. The IO-Q-MPOMDP upper bound is defined by plugging in the true initial state distribution b^0, restricted to factors in c: $\hat{V}_c^P \triangleq V_c^P(\bar{b}_{IO}^0)$.

First we establish a relation between the different vectors representing $V_{I \to c}^P$ and $V_{IO,c}^P$.

Lemma 1. *Let $\pi_c^{t:h-1}$ be a $(h-t)$-steps-to-go policy. Let $\nu \in \mathcal{V}$ and $\nu_{IO} \in \mathcal{V}_{IO}$ be the vectors induced by $\pi_c^{t:h-1}$ under regular MPOMDP back-projections (for some $I_{\to c}$), and under IO back-projections respectively. Then*

$$\forall_{x_c^t} \qquad \max_{D_c^{t+1}} \nu(\langle x_c^t, D_c^{t+1} \rangle) \leq \nu_{IO}(x_c^t).$$

Proof. The proof is listed in Appendix A. □

---

[8]In this section and the next, we will restrict ourselves to rewards of the form $R(s, a)$ to reduce the notational burden, but the presented formulas can be extended to deal with $R(s, a, s')$ formulations in a straightforward way.

[9]For instance, in the case of FFG from Fig. 2, we can imagine an SP that encodes optimistic assumptions by assuming that the neighboring agents always will fight fire at houses i and $i + 2$. Even though it may not be possible to define such a model for all problems—the optimistic influence could depend on the local (belief) state in intricate ways—this gives some interpretation to $\bar{b}_{IO}^t(x_c^t)$. Additionally, we exploit the fact that construction of an optimistic SP model is possible for the considered domains in Section 6.

This lemma provides a strong result on the relation of values computed under regular MPOMDP backups versus influence-optimistic ones. It allows us to establish the following theorem:

Theorem 2. *For an SP c, for all $I_{\to c}$,*

$$\forall_{\bar{b}^t_{I\to c}} \qquad V^P_{I\to c}(\bar{b}^t_{I\to c}) \le V^P_{IO,c}(\bar{b}^t_{IO}),$$

provided that $\bar{b}^t_{IO}$ concides with the marginals of $\bar{b}^t_{I\to c}$:

$$(A1) \quad \forall_{x^t_c} \sum_{D^{t+1}_c} \bar{b}^t_{I\to c}(\langle x^t_c, D^{t+1}_c\rangle) = \bar{b}^t_{IO}(x^t_c).$$

Proof. We start with the left hand side:

$$V^P_{I\to c}(\bar{b}^t_c) = \max_{\nu\in\mathcal{V}} \bar{b}^t_{I\to c}\cdot\nu$$

$$= \max_{\nu\in\mathcal{V}} \sum_{\langle x^t_c, D^{t+1}_c\rangle} \bar{b}^t_{I\to c}(\langle x^t_c, D^{t+1}_c\rangle)\nu(\langle x^t_c, D^{t+1}_c\rangle)$$

$$\le \max_{\nu\in\mathcal{V}} \sum_{x^t_c}\sum_{D^{t+1}_c} \bar{b}^t_{I\to c}(\langle x^t_c, D^{t+1}_c\rangle) \max_{D^{t+1}_c}\nu(\langle x^t_c, D^{t+1}_c\rangle)$$

$$\{A1\} = \max_{\nu\in\mathcal{V}} \sum_{x^t_c} \bar{b}^t_{IO}(x^t_c)\max_{D^{t+1}_c}\nu(\langle x^t_c, D^{t+1}_c\rangle)$$

$$\{Lemma\ 1\}$$

$$\le \max_{\nu_{IO}\in\mathcal{V}_{IO}} \sum_{x^t_c} \bar{b}^t_{IO}(x^t_c)\nu_{IO}(x^t_c)$$

$$= \max_{\nu_{IO}\in\mathcal{V}_{IO}} \bar{b}^t_{IO}\cdot\nu_{IO}$$

$$= V^P_{IO,c}(\bar{b}^t_{IO}),$$

thus proving the theorem. □

Corollary 1. *IO-Q-MPOMDP yields an upper bound to the locally-optimal value: $V^{LO}_c \le \hat{V}^P_c$.*

Proof. The initial beliefs are defined such that the above condition (A1) holds. That is:

$$\sum_{D^{t+1}_c} \bar{b}^0_{I\to c}(\langle x^0_c, D^1_c=\emptyset\rangle) = \bar{b}^0_{IO}(x^t_c) = b^0(x^t_c).$$

Therefore, application of Theorem 2 to the initial belief yields:

$$\forall_{I\to c} \hat{V}^P_c \triangleq V^P_{IO,c}(\bar{b}^0_{IO}) \ge V^P_{I\to c}(\bar{b}^0_{I\to c}) \triangleq V^{MPOMDP}_c(I_{\to c})$$

It is well-known that the MPOMDP value is an upper bound to the Dec-POMDP value [25], such that

$$V^{MPOMDP}_c(I_{\to c}(\pi^{LO}_{\ne c})) \ge V_c(I_{\to c}(\pi^{LO}_{\ne c})),$$

and we can immediately conclude that

$$\hat{V}^P_c \ge V_c(I_{\to c}(\pi^{LO}_{\ne c})) = \max_{\pi_{\ne c}} V_c(I_{\to c}(\pi_{\ne c})) = V^{LO}_c,$$

with the identities given by (5), proving the result. □

4.3 A Dec-POMDP Approach

The previous approaches compute upper bounds by, apart from the IO assumption, additionally making optimistic assumptions on observability or communication capabilities. Here we present a general method for computing *Dec-POMDP-based upper bounds* that, other than the optimistic assumptions about neighboring SPs, make no additional assumptions and thus provide the tightest bounds out of the three that we propose. This approach builds on the recent insight [17, 8, 24] that a Dec-POMDP can be converted to a special case of POMDP (for an overview of this reduction, see [23]); we can thereby leverage the influence-optimistic back-projection (11) to compute an IO-UB that we refer to as *IO-Q-Dec-POMDP*.

As in the previous two sub-sections, we will leverage optimism with respect to an influence-augmented model that we will never need to construct. In particular, as explained in Section 3 we can convert an SP $\mathcal{M}_c$ to an IASP $\mathcal{M}^{IA}_c$ given an influence $I_{\to c}$. Since such an IASP is a Dec-POMDP, we can convert it to a special case of a POMDP:

Definition 6. A *plan-time influence-augmented sub-problem*, $\mathcal{M}^{\text{PT-IA}}_c$, is a tuple $\mathcal{M}^{\text{PT-IA}}_c(\mathcal{M}_c, I_{\to c}) = \langle \check{\mathcal{S}}, \check{\mathcal{A}}, \check{T}_{I\to c}, \check{R}, \check{\mathcal{O}}, \check{O}, \check{h}, \check{b}^0\rangle$, where:

- $\check{\mathcal{S}}$ is the set of states $\check{s}^t = \langle \bar{x}^t_c, \vec{o}^t_c\rangle = \langle x^t_c, D^{t+1}_c, \vec{o}^t_c\rangle$.

- $\check{\mathcal{A}}$ is the set of actions, each $\check{a}^t$ corresponds to a local joint decision rule δ^t_c in the SP.

- $\check{T}_{I\to c}(\check{s}^{t+1}|\check{s}^t, \check{a}^t)$ is the transition function defined below.

- $\check{R}(\check{s}^t, \check{a}^t) = R_c(x^t_c, \delta^t_c(\vec{o}^t_c))$.

- $\check{\mathcal{O}} = \{NULL\}$.

- $\check{O}$ specifies that observation $NULL$ is received with probability 1 (irrespective of the state and action).

- The horizon is not modified: $\check{h} = h$.

- $\check{b}^0$ is the initial state distribution. Since there is only one $\vec{o}^0$ (i.e., the empty joint observation history).

The transition function specifies:

$$\check{T}_{I\to c}(\langle x^{t+1}_c, D^{t+2}_c, \vec{o}^{t+1}_c\rangle \,|\, \langle x^t_c, D^{t+1}_c, \vec{o}^t_c\rangle, \delta^t_c)$$
$$\triangleq \bar{T}_{I\to c}(x^{t+1}_c|\langle x^t_c, D^{t+1}_c\rangle, \delta^t_c(\vec{o}^t_c))\bar{O}(o^{t+1}_c|\delta^t_c(\vec{o}^t_c), x^{t+1}_c)$$

if $\vec{o}^{t+1}_c = (\vec{o}^t_c, o^{t+1}_c)$ and 0 otherwise. In this equation $\bar{T}_{I\to c}, \bar{O}$ are given by the IASP (cf. Section 3).[10]

This reduction shows that it is possible to compute $V^*_c(I_{\to c}(\pi_{\ne c}))$, the optimal value for an SP *given an influence point* $I_{\to c}$, but the formulation is subject to the same computational burden as solving a regular IASP: constructing it is complex due to the inference that needs to be performed to compute $I_{\to c}$, and subsequently solving the IASP is complex due to the large number of augmented states $\check{s}^t = \langle x^t_c, D^{t+1}_c, \vec{o}^t_c\rangle$.

---

[10]Remember that D^{t+2}_c is a function of the specified quantities: $D^{t+2}_c = d(x^t_c, D^{t+1}_c, \delta^t_c(\vec{o}^t_c), x^{t+1}_c)$.

Fortunately, here too we can compute an upper bound to *any* feasible incoming influence, and thus to V_c^{LO}, by using optimistic backup operations with respect to an underspecified model, to which we refer as simply *plan-time SP*:

Definition 7. We define the *plan-time sub-problem* $\mathcal{M}_c^{PT}$ as an under-specified POMDP $\mathcal{M}_c^{\mathrm{PT}}(\mathcal{M}_c,\cdot) = \left\langle \check{\mathcal{S}},\check{\mathcal{A}},\check{T}_{(\cdot)},\check{R},\check{\mathcal{O}},\check{O},\check{h},\check{b}^0 \right\rangle$ with

- states of the form $\check{s}^t = \left\langle x_c^t,\vec{o}_c^t \right\rangle$,
- an underspecified transition model

$$\check{T}_{(\cdot)}(\check{s}^{t+1}|\check{s}^t,\check{a}^t) \triangleq$$
$$T_c(x_c^{t+1}|x_c^t,\delta_c^t(\vec{o}_c^t),u_c^{t+1})O_c(o_c^{t+1}|\delta_c^t(\vec{o}_c^t),x_c^{t+1}),$$

- and $\check{\mathcal{A}},\check{R},\check{\mathcal{O}},\check{O},\check{h},\check{b}^0$ as above.

Since this model is a special case of a POMDP, the theory developed in Section 4.2 applies: we can maintain a plan-time sufficient statistic $\sigma^{c,t}$ (essentially the 'belief' $\check{b}$ over augmented states $\check{s}^t = \left\langle x_c^t,\vec{o}_c^t \right\rangle$) and we can write down the value function using (9). Most importantly, the IO back-projection (11) also applies, which means that (similar to the MPOMDP case) we can avoid ever constructing the full PT-IASP. The IO back-projection in this case translates to:

$$\nu_{IO}^{\delta_c^t}(x_c^t,\vec{o}_c^t) \triangleq \max_{u_c^{t+1}} \sum_{x_c^{t+1}} \Pr(o_c^{t+1}|\delta_c^t(\vec{o}_c^t),x_c^{t+1})$$
$$\Pr(xn_c^{t+1}|x_c^t,\delta_c^t(\vec{o}_c^t),u_c^{t+1})\Pr(xl_c^{t+1}|x_c^t,\delta_c^t(\vec{o}_c^t))$$
$$\nu_{IO}(x_c^{t+1},\vec{o}_c^{t+1}). \quad (12)$$

Here, we omitted the superscript for the *NULL* observation. Also, note that $O(o_i^{t+1}|a_i^t,x_i^{t+1})$ in (11) corresponds to the *NULL* observation in the PT model, but since the observation histories are in the states, $\Pr(o_c^{t+1}|\delta_c^t(\vec{o}_c^t),x_c^{t+1})$ comes out of the *transition* model.

Again, given this modified back-projection, the IO-Q-Dec-POMDP value $\hat{V}_c^D$ can be computed using any *exact* POMDP solution method that makes use of vector back-projections; all that is required is to substitute these back projections by their modified form (12).

Corollary 2. *IO-Q-Dec-POMDP yields an upper bound to the locally-optimal value:* $V_c^{LO} \leq \hat{V}_c^D$.

Proof. Directly by applying Corollary 1 to $\mathcal{M}_c^{\mathrm{PT}}$. $\square$

4.4 Computational Complexity

Due to the maximization in (6), (11) and (12), IO back-projections are more costly than regular (non-IO) back-projections. Here we analyze the computational complexity of the proposed algorithms relative to the regular, non-IO, backups.

We start by comparing IQ-Q-MMDP backup operation (6) to the regular MMDP backup for an SP that does not have incoming influences. For such an SP, the

MMDP backup is given by

$$Q(x_c^t,a_c^t) = \sum_{x_c^{t+1}} \Pr(x_c^{t+1}|x_c^t,a_c^t)$$
$$\left[R_c(x_c^t,a_c^t,x_c^{t+1}) + \max_{a_c^{t+1}} Q(x_c^{t+1},a_c^{t+1}) \right]. \quad (13)$$

Comparing (6) and (13), we see two differences: 1) the transition probabilities can be written in one term since we do not need to discriminate NLAFs from OLAFs, and 2) there is no maximization over influence sources u_c^{t+1}. The first difference does not induce a change in computational cost, only in notation: in both cases the entire transition probability is given as the product of next-stage CPTs. The second difference does induce a change in computational cost: in (6), in order to select the maximum, the inner part of the right-hand side needs to be evaluated for each instantiation of the influence sources $\left|u_c^{t+1}\right|$. That is, the computational complexity of a Q-MMDP backup (for a particular (x_c^t,a_c^t)-pair) is

$$O(|\mathcal{X}_c| \times |a_c|),$$

whereas the total computational cost of the IO-Q-MMDP backup is:

$$O(\left|u_c^{t+1}\right| \times |\mathcal{X}_c| \times |a_c|).$$

As such, the complexity of each backup is multiplied by the number of influence source instantiations $\left|u_c^{t+1}\right|$. This means that the overhead of IO-Q-MMDP relative to Q-MMDP is $O(\left|u_c^{t+1}\right|)$.

A similar comparison of the overhead of the IO-Q-MPOMDP back projection (11) with respect to the regular Q-MPOMDP back projection (10) leads to exactly the same relative overhead of $O(\left|u_c^{t+1}\right|)$. Since IO-Q-Dec-POMDP makes use of this last result, also in this case the relative overhead is $O(\left|u_c^{t+1}\right|)$.

Concluding, the computational complexity of the methods we proposed to compute IO-UBs is given by multiplying the computational complexity of the 'underlying' MMDP, MPOMDP or Dec-POMDP method with $\left|u_c^{t+1}\right|$. This means that the relative overhead, is equal for all methods. Since the amount of overhead is a linear function of $\left|u_c^{t+1}\right|$, it is problem dependent: more densely coupled problems will lead to a higher overhead, while very loosely coupled problems have a low overhead.

5. GLOBAL UPPER BOUNDS

We next discuss how the methods to compute local upper bounds can be employed in order to compute a global upper bound for factored Dec-POMDPs.

The basic idea is to apply a non-overlapping decomposition $\mathcal{C}$ (i.e., a partitioning) of the reward functions $\left\{R^l\right\}$ of the original factored Dec-POMDP into SPs $c \in \mathcal{C}$, and to compute an IO upper bound $\hat{V}_c^{IO}$ for each (which can be any of the three IO-UBs proposed in Section 4). Our *global influence-optimistic upper bound* is then given by:

$$\hat{V}^{IO} \triangleq \sum_{c \in \mathcal{C}} \hat{V}_c^{IO}. \quad (14)$$
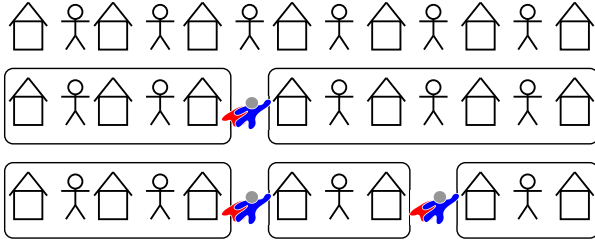
Figure 3: Illustration of construction of a global upper bound for 6-agent FFG using influence-optimistic upper bounds for sub-problems.

Fig. 3 illustrates the construction of a global upper bound $\hat{V}$ for the 6-agent FFG, showing the original problem (top row) and two possible decompositions in SPs. The second row specifies a decomposition into two SPs, while the third row uses three SPs. The illustration clearly shows how (in this problem) a decomposition eliminates certain agents completely and replaces them with optimistic assumptions: E.g., in the second row, during the computation of $\hat{V}_c^{IO}$ for both SPs ($c = 1,2$) the assumption is made that agent 3 will always fight fire in the SP under concern. Effectively we assume that agent 3 fights fire at both house 3 and house 4 simultaneously (and hence is represented by a superhero figure). Fig. 3 also illustrates that, due to the line structure of FFG, there are two *types* of SPs: 'internal' SPs which make optimistic assumptions on two sides, and 'edge' SPs that are optimistic at just one side.

Finally, we formally prove the correctness of our proposed upper bounding scheme.

Theorem 3. *Let $\mathcal{C}$ be a partitioning of the reward function set $\mathcal{R}$ into sub-problems such that every R^l is represented in one SP c, then the global IO-UB is in fact an upper bound to the optimal value $\hat{V}^{IO} \geq V(\pi^*)$.*

Proof. Starting from the definition (14), we have

$$
\begin{aligned}
\hat{V}^{IO} &\triangleq \sum_{c \in \mathcal{C}} \hat{V}_c^{IO} \overset{\{\text{Section } 4\}}{\geq} \sum_{c \in \mathcal{C}} V_c^{LO} \\
&\triangleq \sum_{c \in \mathcal{C}} \max_{\pi_{\neq c}} V_c^{BR}(\pi_{\neq c}) \\
&\triangleq \sum_{c \in \mathcal{C}} \max_{\pi_{\neq c}} \max_{\pi_c} V_c(\pi_c, \pi_{\neq c}) \\
&\geq \max_{\pi} \sum_{c \in \mathcal{C}} V_c(\pi_c, \pi_{\neq c}) \overset{\{\mathcal{C}\text{is a partition}\}}{=} V(\pi^*)
\end{aligned}
$$

thus proving the result. □

6. EMPIRICAL EVALUATION

In order to test the potential impact of the proposed influence-optimistic upper bounds, we present numerical results in the context of a number of benchmark problems. In this evaluation, we focus on the (relative) values found by these heuristics, as we hope that these will

spark a number of interesting ideas for further research (such as the notion of 'influence strength', its relation to approximability of factored Dec-POMDPs, and the key idea that reasonable bounds for very large problems may be possible). We do not thoroughly investigate timing results as the analysis of Section 4.4 indicates that relative timing results follow those of regular (non-IO) MMDP, MPOMDP and Dec-POMDP methods; see, e.g., [25] for a comparison of such timing results. However, in order to provide a overall idea of the run times, we do provide some indicative running times.[11]

6.1 Comparison of Different Bounds

The bounds that we propose are ordered in tightness, $\hat{V}_c^D \leq \hat{V}_c^P \leq \hat{V}_c^M$, similar to how regular (non-IO) Dec-POMDP, Q-MPOMDP, and Q-MMDP values relate [25]. To get an understanding of how these differences turn out in practice, Fig. 4(left) compares the different upper bounds introduced.

Although the approach described in the paper is general, in the numerical evaluation here we exploit the property that the optimistic influences are easily identified off-line, which allows for the construction of small 'optimistic Dec-POMDPs' (respectively MPOMDPs or MMDPs) without sacrificing in bound quality. E.g., in order to compute the local IQ-Q-Dec-POMDP upper bound for a 3-house FFG 'edge' SP, we define a regular 3-house Dec-POMDP where the transitions probabilities for the first house (say X^i in Fig. 2) are modified to account for the optimistic assumption that another (superhero) agent fights fire there and that its neighbor is not burning (i.e., $a_{i-1} = right$ and $X^{i-1} = not\_burning$ in Fig. 2). The values $\hat{V}_c^D$ (resp. $\hat{V}_c^P, \hat{V}_c^M$) are computed by running a state-of-the-art Dec-POMDP solver [24] (resp. incremental pruning [6], plain dynamic programming) on such optimistically defined problems.

Fig. 4(left) shows the values for such 'edge' problems. Missing bars indicate time-outs (>4h). As an indication of run time, the $|\mathcal{D}'| = 3, h = 5$ problem took $2.21s$ for IO-Q-MMDP, and $995.28s$ for IO-Q-Dec-POMDP. The shown values indicate that $\hat{V}_c^D, \hat{V}_c^P$ can be tighter than $\hat{V}_c^M$ in practice. In most cases, the difference between IO-Q-MPOMDP and IO-Q-Dec-POMDP is small, but these could become larger for longer horizons [25]. We performed the same analysis for the ALOHA benchmark [27], and found very similar results.

We also compare the bounds found on the different types of SPs (internal and edge-cases, see Fig. 3) encountered in FFG ($h = 4$). In addition, Fig. 4(middle) also includes—if computable within the allowed time—values of SPs that are 'full' problems (i.e., the regular optimal Dec-POMDP value for the full FFG instance with the indicated number of agents.) This makes clear that the optimistic assumption has quite some effect: being optimistic at one edge more than halves the optimal cost, and the IO assumption at both edges of the SP leads to another significant reduction of that cost. This

---

[11] All experiments are run on a Intel Xeon E5-2650L, 32GB system making use of one core only.
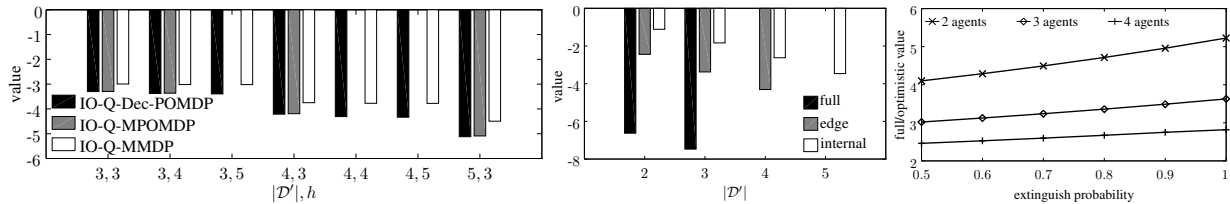
Figure 4: Left: Different IO-UBs on 'edge' FFG problems. Middle: IO-Q-Dec-POMDP upper bound on different types (internal, edge, full) of SPs. Right: The impact the 'influence strength'.

is to be expected: the optimistic problems assume that there *always* will be another agent fighting fire at the house at an optimistic edge, while the full problem *never* has another agent at that same house. When also taking into account the transition probabilities—two agents at a house will completely extinguish a fire—it is clear that the IO assumption should have a high impact on the local value.

6.2 The Effect of Influence Strength

Fig. 4(middle) makes clear that the IO assumption in FFG is quite strong and leads to a significant over-estimation of the local value compared to the 'full' problem. We say that FFG has a high *influence strength.* In fact, this hints at a new dimension of the qualification of weak coupling [44] that takes into account the variance in NLAF probabilities (as a function of the change of value of the influence source) and their impact on the local value.

As a preliminary investigation of this concept, we devise a modification of FFG where the influence strength can be controlled. In particular, we parameterize the probability that a fire is extinguished completely when 2 agents visit the same house, which is set to 1 in the original problem definition. Lower values of this probability mean that optimistically assuming there is another agent at a house will lead to less advantage, and thus lower influence strength.

Fig. 4(right) shows the results of this experiment. It shows that there is a clear relation between the fire-extinguish probability when two agents fight fire at a house, and the ratio between the 'regular' (Dec-POMDP) value and optimistic value. It also shows that SPs with more agents are less affected: this makes sense since optimistic assumptions account for a smaller fraction of the achievable value. In other words, larger sub-problems give a tighter approximation.

6.3 Bounding Heuristic Methods

Here we investigate the ability to provide informative *global* upper bounds. While the previous analysis shows that the overestimation is quite significant at the *true* edges of the problem (where no agents exist), this is not necessarily informative of the overestimation at internal edges in decompositions of larger problems (where other agents do exist, even if not superheros). As such, besides investigating the upper bounding capability, the analysis here also provides a better understanding of such
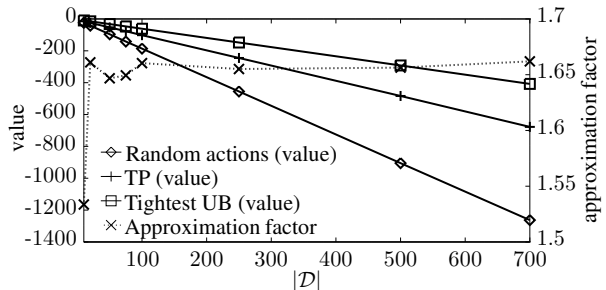

Figure 5: The global IO-Q-Dec-POMDP upper bound on large FFG instances.

internal overestimations.

We use the tightest upper bound we could find by considering different SP partitions, with sizes ranging from $|\mathcal{D}| = 2$–5, and investigate the guarantees that it can provide for transfer planning (TP) [27], which is one of the methods capable of providing solutions for large factored Dec-POMDPs. Since the method is a heuristic method that does not provide the exact value of the reported joint policy, the value of TP, V^{TP}, is determined using 10.000 simulations of the found joint policy leading to accurate estimates.[12] To put the results into context, we also show the value of a random policy. Finally, we show (second y-axis in Fig. 5) what we call the *empirical approximation factor (EAF)*:

$$EAF = \max\{\frac{\hat{V}^{IO}}{V^{TP}}, \frac{V^{TP}}{\hat{V}^{IO}}\}.$$

This is a number comparable to the approximation factors of *approximation algorithms* [39].[13]

We computed upper bounds for large, horizon $h = 4$, FFG instances. The computation of the local upper bounds for the largest SPs used (i.e., $|\mathcal{D}'| = 5$) took 3.31 secs for IO-Q-MMDP and 2696.23s for IO-Q-Dec-POMDP. Fig. 5 shows the results that indicate that the upper bound is relatively tight: the solutions found by TP are not too far from the upper bound. In particular, the EAF lies typically between 1.4 and 1.7, thus

---

[12]Note that there is no method for Dec-POMDP policy evaluation that runs in polynomial time. In fact, existence of such a method would reduce the complexity of solving a Dec-POMDP to NP, an impossibility since the time hierarchy theorem implies that NP$\neq$ NEXP.

[13]'Empirical' emphasizes the lack of *a priori* guarantees.

$\|\mathcal{D}\|$	50	75	100	250
V^{TP}	-71.99	-111.07	-148.70	-382.47
$\hat{V}^{IO}$	-72.00	-107.06	-144.00	-360.00
EAF	1.00	1.04	1.03	1.06

Table 1: Empirical approximation factors for Aloha ($h = 3$) with varying number of agents.

		h=2		h=3		h=4		h=5
		HOUSESEARCH : *Diamond* : Pr(correct obs.) = 0.75						
	#	s	#	s	#	s	#	s
M	55	2.36	245	33.86	2072	1179		
P	55	2.42	241	34.21	1744	1010		
M'	45	2.00	91	9.00	183	64.32	97	991.2
P'	45	2.02	91	10.05	119	39.84	66	607.9
		HOUSESEARCH : *Diamond* : Pr(correct obs.) = 0.5						
	#	s	#	s	#	s	#	s
M	47	3.35	225	51.80	1434	1935		
P	47	3.32	223	52.75	1416	1905		
M'	23	1.96	24	4.69	41	38.42	148	1882
P'	11	0.89	15	3.90	19	25.32	23	256.8
		HOUSESEARCH : *Squares* : Pr(correct obs.) = 0.75						
	#	s	#	s	#	s		
M	25	2.41	311	120.5	1323	5846		
P	25	2.56	311	121.3	1032	7184		
M'	25	2.61	277	109.6	816	3924		
P'	25	2.53	240	95.28	752	6760		

Table 2: Node count and run time for A*-OIS.

providing firm guarantees for solutions of factored Dec-POMDPs with up to 700 agents. Moreover, we see that the EAF stays roughly constant for the larger problem instances indicating that *relative* guarantees do not degrade as the number of agents increase. Of course, the question of whether the optimal value lies closer to the blue (UB) or orange (TP) line remains open; only further research on improved (heuristic) solution methods and tighter upper bounds can answer that question. However, we have gone from a situation where the only upper bound we had was 'predict R_{max} for every stage' (which corresponds to the value 0 and EAF=∞) to a situation where we have a much more informative bound.

Results obtained for a similar approach for ALOHA using SPs containing up to $|\mathcal{D}'| = 6$ agents are shown in Table 1. The numbers clearly illustrate that it is possible to provide very strong guarantees for problems up to $|\mathcal{D}| = 250$ agents (beyond which memory forms the bottleneck for TP); the solution for the $|\mathcal{D}| = 50$ instance is essentially optimal, indicating also a very tight bound for this problem.

6.4 Improved Heuristic Influence Search

Aside from analyzing the solution quality of approximate methods, our bounds can also be used in optimal methods. In particular, A*-OIS [41] solves TD-POMDPs (a sub-class of factored Dec-POMDPs) by decomposing them into 1-agent SPs, searching through the space of influences, and pruning using optimistic heuristics. However, existing A*-OIS heuristics treat the unspecified-influence stages of the SPs as fully-observable. In contrast, IO-Q-MPOMDP models the *partial observability* of the SPs.

We now present results that suggest an added computational benefit to treating partial observability in the A*-OIS heuristics. Table 2 illustrates the differences in pruning afforded by four different A*-OIS heuristics: M is the baseline MDP-based heuristic from [41], P is shorthand for IO-Q-MPOMDP, and M' and P' are variations that improve tightness with locally-derived probability bounds on the optimistic influences. We report node counts and runtime across several problem instances from the HouseSearch domain [41].

As shown, the POMDP-based heuristics tends to allow for more pruning (fewer expanded nodes) and thereby reduced computation in comparison to their MDP-based counterparts. Contrasting this reduction across two variants of *Diamond* HouseSearch suggests that the IO-Q-MPOMDP heuristics gain more advantage when observability is more restricted. However, this advantage is

sometimes outweighed by the increased computational overhead of a more complex heuristic calculation (such as in *Squares* HouseSearch $h = 4$). The question of predicting in advance when this will be the case is interesting, but difficult (and orthogonal to our contribution).

7. RELATED WORK

Here we provide an overview of related approaches.

Scalable Heuristic Methods.

In recent years, many scalable approaches without guarantees have been developed for Dec-POMDPs and related models [14, 46, 48, 16, 40, 37, 27, 47, 36]. The upper bounding mechanism that we propose could be useful for benchmarking many of these methods.

Sub-Problems vs. Source Problems.

The use of sub-problems (SPs) is conceptually similar to the use of *source problems* in transfer planning (TP) [27]. Differences are that, in our partitioning-based upper bounding scheme, SPs are selected such that they do not contain overlapping rewards, while TP allows for overlapping source problems. Moreover TP does not consider optimistic influences but implicitly assumes arbitrary influences. Finally, TP is used as a way to compute a heuristic for the original problem; our approach here simply returns a scalar value (although extensions to use IO heuristics for heuristic search for factored Dec-POMDPs are an interesting direction for future research).

Optimism with Respect to Influences.

Being optimistic with respect to external influences has been considered before. For instance, Kumar & Zilberstein [14] make optimistic assumptions on transitions in an ND-POMDP to derive an MMDP-based policy which is used to sample belief points for memory-

bounded dynamic programming. The approach does not use these assumptions to upper bound the global value and the formulation is specific to ND-POMDPs. As described in the experiments, Witwicki, Oliehoek & Kaelbling [41] use a local upper bound in order to perform heuristic influence search for TD-POMDPs. IO-Q-MMDP can be seen as a generalization of that heuristic to both factored Dec-POMDPs and multiagent sub-problems, while our other heuristics additionally deal with partial observability.

Quality Guarantees for Large Dec-POMDPs.

As mentioned in the introduction, a few approaches to computing upper bound for large-scale MASs and employing them in heuristic search methods have been proposed. In particular, there are some scalable approaches with guarantees for the case of transition and observation independence as encountered in TOI-MDPs and ND-POMDPs [3, 2, 38, 19, 10, 9]. The approach by Witwicki for TD-POMDPs [42, Sect. 6.6] is a bit more general in that it allows some forms of transition dependence, as long as the interactions are *directed* (one agent can affect another) and no two agents can affect the same factor in the same stage. In addition, scalability relies on each agent having only a handful of 'interaction ancestors'.

However, these previous approaches rely on the true value function being factored as the sum of a set $\mathcal{E}$ of *local* value components:

$$V(\pi) = \sum_{e \in \mathcal{E}} V_e(\pi_e),$$

where π_e is the *local* joint policy of the agents that participate in component e. (This is also referred to as the 'value factorization' framework [16]). For this setting, an upper bound is easily constructed as the sum of local upper bounds:

$$\hat{V}(\pi) = \sum_{e \in \mathcal{E}} \hat{V}_e(\pi_e).$$

While this resembles our IO upper bound (14), the crucial distinction is that for value factorized settings computing $\hat{V}_e$ does not require any influence-optimism: the reason that value-factorization holds is precisely *because there are no influence sources* for the components. As such, our influence-optimistic upper bounds can be seen as a strict generalization of the upper bounds that have been employed for settings with factored value functions. Investigating if such methods, such as the method by Dibangoye et al. [9] can be modified to use our IO-UBs is an interesting direction of future work.

Finally, the event-detecting multiagent MDP [15] provides quality guarantees for a specific class of sensor network problems by using the theory of submodular function maximization. It is the only previous method with quality guarantees that delivers scalability with respect to the number of agents without assuming that the value function of the problem is additively factored into small components.

8. CONCLUSIONS

We presented a family of *influence-optimistic* upper bounds for the value of sub-problems of factored Dec-POMDPs, together with a partition-based decomposition approach that enables the computation of global upper bounds for very large problems. The approach builds upon the framework of influence-based abstraction [28], but—in contrast to that work—makes optimistic assumptions on the incoming 'influences', which makes the sub-problems easier to solve. An empirical evaluation compares the proposed upper bounds and demonstrates that it is possible to achieve guarantees for problems with hundreds of agents, showing that found heuristic solution are in fact close to optimal (empirical approximation factors of < 1.7 in all cases and sometimes substantially better). This is a significant contribution, given the complexity of computing ϵ-approximate solutions and the fact that tight global upper bounds are of crucial importance to interpret the quality of heuristic solutions.

Intuitively, the proposed approach is expected to work well in settings where the Dec-POMDP is 'weakly' coupled. The work by Witwicki & Durfee [44] identifies three dimensions that can be used to quantify the notion of weak coupling. Our experiments suggest the existence of a new dimension that can be thought of as *influence strength*. This dimension captures the impact of non-local behavior on local *values* and thus directly relates to how well a problem can be approximated using localized components.

In this paper we focused on the finite-horizon case, but the principle of influence optimism underlying the upper-bounding approach can also be applied in infinite-horizon settings. Furthermore, the approach can be modified to compute 'pessimistic' influence (i.e., lower) bounds, which could be useful in competitive settings, or for risk-sensitive planning [20]. It is also immediately applicable to problems involving 'unpredictable' dynamics [45, 7]. Finally, our upper-bounding method contributes a useful precursor for techniques that automatically search the space of possible upper bounds decompositions, efficient optimal influence-space heuristic search methods (for which we provided preliminary evidence in this paper), and A* methods for a large class of factored Dec-POMDPs. In particular, a promising idea is to employ our factored upper bounds in combination with the heuristic search methods by Dibangoye et al. [9]. While it is not possible to directly use that method since it additionally requires a factored lower bound function, pessimistic-influence bounds could provide those.

A limitation of the current approach is that sub-problems still need to be relatively small, since we rely on *optimal* optimistic solution of the sub-problems. Developing more scalable 'optimistic solution methods' thus is an important direction of future work. Experiments with influence search indicate that using probabilistic bounds on the positive influences has a major impact [41]. As such, another important direction of future work

is investigating if it is possible to develop tighter upper bounds by making more realistic optimistic assumptions.

9. REFERENCES

[1] C. Amato and F. A. Oliehoek. Bayesian reinforcement learning for multiagent systems with state uncertainty. In *AAMAS Workshop on Multi-Agent Sequential Decision Making in Uncertain Domains (MSDM)*, pages 76–83, 2013.

[2] R. Becker, S. Zilberstein, and V. Lesser. Decentralized Markov decision processes with event-driven interactions. In *Proc. of the International Conference on Autonomous Agents and Multiagent Systems*, pages 302–309, 2004.

[3] R. Becker, S. Zilberstein, V. Lesser, and C. V. Goldman. Transition-independent decentralized Markov decision processes. In *Proc. of the International Conference on Autonomous Agents and Multiagent Systems*, pages 41–48, 2003.

[4] C. Boutilier, T. Dean, and S. Hanks. Decision-theoretic planning: Structural assumptions and computational leverage. *Journal of Artificial Intelligence Research*, 11:1–94, 1999.

[5] C. Boutilier and D. Poole. Computing optimal policies for partially observable decision processes using compact representations. In *Proc. of the National Conference on Artificial Intelligence*, pages 1168–1175, 1996.

[6] A. Cassandra, M. L. Littman, and N. L. Zhang. Incremental pruning: A simple, fast, exact method for partially observable Markov decision processes. In *Proc. of Uncertainty in Artificial Intelligence*, pages 54–61, 1997.

[7] K. V. Delgado, S. Sanner, and L. N. De Barros. Efficient solutions to factored MDPs with imprecise transition probabilities. *Artificial Intelligence*, 175(9):1498–1527, 2011.

[8] J. S. Dibangoye, C. Amato, O. Buffet, and F. Charpillet. Optimally solving Dec-POMDPs as continuous-state MDPs. In *Proc. of the International Joint Conference on Artificial Intelligence*, 2013.

[9] J. S. Dibangoye, C. Amato, O. Buffet, and F. Charpillet. Exploiting separability in multiagent planning with continuous-state MDPs. In *Proc. of the International Conference on Autonomous Agents and Multiagent Systems*, pages 1281–1288, 2014.

[10] J. S. Dibangoye, C. Amato, A. Doniec, and F. Charpillet. Producing efficient error-bounded solutions for transition independent decentralized MDPs. In *Proc. of the International Conference on Autonomous Agents and Multiagent Systems*, pages 539–546, 2013.

[11] R. Emery-Montemerlo, G. Gordon, J. Schneider, and S. Thrun. Approximate solutions for partially observable stochastic games with common payoffs. In *Proc. of the International Conference on Autonomous Agents and Multiagent Systems*, pages 136–143, 2004.

[12] E. A. Hansen, D. S. Bernstein, and S. Zilberstein. Dynamic programming for partially observable stochastic games. In *Proc. of the National Conference on Artificial Intelligence*, pages 709–715, 2004.

[13] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2):99–134, 1998.

[14] A. Kumar and S. Zilberstein. Constraint-based dynamic programming for decentralized POMDPs with structured interactions. In *Proc. of the International Conference on Autonomous Agents and Multiagent Systems*, pages 561–568, 2009.

[15] A. Kumar and S. Zilberstein. Event-detecting multi-agent MDPs: Complexity and constant-factor approximations. In *Proc. of the International Joint Conference on Artificial Intelligence*, pages 201–207, 2009.

[16] A. Kumar, S. Zilberstein, and M. Toussaint. Scalable multiagent planning using probabilistic inference. In *Proc. of the International Joint Conference on Artificial Intelligence*, pages 2140–2146, 2011.

[17] L. C. MacDermed and C. Isbell. Point based value iteration with optimal belief compression for Dec-POMDPs. In *Advances in Neural Information Processing Systems 26*, pages 100–108. 2013.

[18] A. Mahajan and M. Mannan. Decentralized stochastic control. *Annals of Operations Research*, 2014.

[19] J. Marecki, T. Gupta, P. Varakantham, M. Tambe, and M. Yokoo. Not all agents are equal: scaling up distributed POMDPs for agent networks. In *Proc. of the International Conference on Autonomous Agents and Multiagent Systems*, pages 485–492, 2008.

[20] J. Marecki and P. Varakantham. Risk-sensitive planning in partially observable environments. In *Proc. of the International Conference on Autonomous Agents and Multiagent Systems*, pages 1357–1368, 2010.

[21] J. V. Messias, M. T. J. Spaan, and P. U. Lima. Efficient offline communication policies for factored multiagent POMDPs. In *NIPS 24*, pages 1917–1925. 2011.

[22] R. Nair, P. Varakantham, M. Tambe, and M. Yokoo. Networked distributed POMDPs: A synthesis of distributed constraint optimization and POMDPs. In *Proc. of the National Conference on Artificial Intelligence*, pages 133–139, 2005.

[23] F. A. Oliehoek and C. Amato. Dec-POMDPs as non-observable MDPs. Technical report, University of Amsterdam, 2014.

[24] F. A. Oliehoek, M. T. J. Spaan, C. Amato, and S. Whiteson. Incremental clustering and expansion for faster optimal planning in decentralized POMDPs. *Journal of AI Research*, 46:449–509, 2013.

[25] F. A. Oliehoek, M. T. J. Spaan, and N. Vlassis. Optimal and approximate Q-value functions for decentralized POMDPs. *Journal of AI Research*, 32:289–353, 2008.

[26] F. A. Oliehoek, M. T. J. Spaan, S. Whiteson, and N. Vlassis. Exploiting locality of interaction in factored Dec-POMDPs. In *Proc. of the International Conference on Autonomous Agents and Multiagent Systems*, pages 517–524, 2008.

[27] F. A. Oliehoek, S. Whiteson, and M. T. J. Spaan. Approximate solutions for factored Dec-POMDPs with many agents. In *Proc. of the International Conference on Autonomous Agents and Multiagent Systems*, pages 563–570, 2013.

[28] F. A. Oliehoek, S. Witwicki, and L. P. Kaelbling. Influence-based abstraction for multiagent systems. In *Proc. of the AAAI Conference on Artificial Intelligence*, pages 1422–1428, 2012.

[29] J. M. Ooi and G. W. Wornell. Decentralized control of a multiple access broadcast channel: Performance bounds. In *Proc. of the 35th Conference on Decision and Control*, pages 293–298, 1996.

[30] J. Pajarinen and J. Peltonen. Efficient planning for factored infinite-horizon DEC-POMDPs. In *Proc. of the International Joint Conference on Artificial Intelligence*, pages 325–331, 2011.

[31] Z. Rabinovich, C. V. Goldman, and J. S. Rosenschein. The complexity of multiagent systems: the price of silence. In *Proc. of the International Conference on Autonomous Agents and Multiagent Systems*, pages 1102–1103, 2003.

[32] M. Roth, R. Simmons, and M. Veloso. Reasoning about joint beliefs for execution-time communication decisions. In *Proc. of the International Conference on Autonomous Agents and Multiagent Systems*, pages 786–793, 2005.

[33] G. Shani, J. Pineau, and R. Kaplow. A survey of point-based POMDP solvers. *Autonomous Agents and Multi-Agent Systems*, 27(1):1–51, 2013.

[34] M. T. J. Spaan. Partially observable Markov decision processes. In *Reinforcement Learning: State of the Art*, pages 387–414. Springer Verlag, 2012.

[35] D. Szer, F. Charpillet, and S. Zilberstein. MAA*: A heuristic search algorithm for solving decentralized POMDPs. In *Proc. of Uncertainty in Artificial Intelligence*, pages 576–583, 2005.

[36] P. Varakantham, Y. Adulyasak, and P. Jaillet. Decentralized stochastic planning with anonymity in interactions. In *Proc. of the AAAI Conference on Artificial Intelligence*, pages 2505–2512, 2014.

[37] P. Varakantham, S. Cheng, G. J. Gordon, and A. Ahmed. Decision support for agent populations in uncertain and congested environments. In *Proc. of the AAAI Conference on Artificial Intelligence*, 2012.

[38] P. Varakantham, J. Marecki, Y. Yabu, M. Tambe, and M. Yokoo. Letting loose a SPIDER on a network of POMDPs: Generating quality guaranteed policies. In *Proc. of the International Conference on Autonomous Agents and Multiagent Systems*, 2007.

[39] V. V. Vazirani. *Approximation Algorithms*. Springer-Verlag, 2001.

[40] P. Velagapudi, P. Varakantham, P. Scerri, and K. Sycara. Distributed model shaping for scaling to decentralized POMDPs with hundreds of agents. In *Proc. of the International Conference on Autonomous Agents and Multiagent Systems*, 2011.

[41] S. Witwicki, F. A. Oliehoek, and L. P. Kaelbling. Heuristic search of multiagent influence space. In *Proc. of the International Conference on Autonomous Agents and Multiagent Systems*, pages 973–981, 2012.

[42] S. J. Witwicki. *Abstracting Influences for Efficient Multiagent Coordination Under Uncertainty*. PhD thesis, University of Michigan, 2011.

[43] S. J. Witwicki and E. H. Durfee. Influence-based policy abstraction for weakly-coupled Dec-POMDPs. In *Proc. of the International Conference on Automated Planning and Scheduling*, pages 185–192, 2010.

[44] S. J. Witwicki and E. H. Durfee. Towards a unifying characterization for quantifying weak coupling in Dec-POMDPs. In *Proc. of the International Conference on Autonomous Agents and Multiagent Systems*, pages 29–36, 2011.

[45] S. J. Witwicki, F. S. Melo, J. C. Fernández, and M. T. J. Spaan. A flexible approach to modeling unpredictable events in MDPs. In *Proc. of the International Conference on Automated Planning and Scheduling*, pages 260–268, 2013.

[46] F. Wu, S. Zilberstein, and X. Chen. Trial-based dynamic programming for multi-agent planning. In *Proc. of the AAAI Conference on Artificial Intelligence*, pages 908–914, 2010.

[47] F. Wu, S. Zilberstein, and N. R. Jennings. Monte-Carlo expectation maximization for decentralized POMDPs. In *Proc. of the International Joint Conference on Artificial Intelligence*, pages 397–403, 2013.

[48] Z. Yin and M. Tambe. Continuous time planning for multiagent teams with temporal constraints. In *Proc. of the International Joint Conference on Artificial Intelligence*, 2011.

APPENDIX

A. PROOFS

Lemma 1. Let $\pi_c^{t:h-1}$ be a $(h-t)$-steps-to-go policy. Let $\nu \in \mathcal{V}$ and $\nu_{IO} \in \mathcal{V}_{IO}$ be the vectors induced by $\pi_c^{t:h-1}$ under regular MPOMDP back-projections (for some $I_{\rightarrow c}$), and under IO back-projections. Then

$$\forall_{x_c^t} \quad \max_{D_c^{t+1}} \nu(\langle x_c^t, D_c^{t+1} \rangle) \leq \nu_{IO}(x_c^t).$$

Proof. The proof is via induction. The base case is for the last stage $t = h-1$, in which case the vectors only consist of the immediate reward:

$$\max_{D_c^{t+1}} \nu(\langle x_c^t, D_c^{t+1} \rangle) = \max_{D_c^{t+1}} R_{a^t}(x_c^t) = R_{a^t}(x_c^t) = \nu_{IO}(x_c^t),$$

where a^t is the joint action specified by $\pi_c^{h-1:h-1}$, thus proving the base case. The induction step follows.

Induction Hypothesis: Suppose that, given that $\nu \in \mathcal{V}$ and $\nu_{IO} \in \mathcal{V}_{IO}$ are vectors for the same policy $\pi_c^{t+1:h-1}$, for all x_c^{t+1}

$$\max_{D_c^{t+2}} \nu(\langle x_c^{t+1}, D_c^{t+2} \rangle) \leq \nu_{IO}(x_c^{t+1}) \tag{15}$$

holds.

To prove: Given that $\nu \in \mathcal{V}$ and $\nu_{IO} \in \mathcal{V}_{IO}$ are vectors for the same policy $\pi_c^{t:h-1}$, for all x_c^t

$$\max_{D_c^{t+1}} \nu(\langle x_c^t, D_c^{t+1} \rangle) \leq \nu_{IO}(x_c^t). \tag{16}$$

Proof:
We first define the vectors from the l.h.s. Let a^t denote the first joint action specified by $\pi_c^{t:h-1}$. Then we can write

$$\nu(\langle x_c^t, D_c^{t+1} \rangle) = r_a(x_c^t) + \gamma \sum_{o^{t+1}} \nu^{\pi|ao}(\langle x_c^t, D_c^{t+1} \rangle) \tag{17}$$

where $\nu^{\pi|ao}$ is the back-projection of the vector $\Gamma(\pi_c^{t:h-1}, a_c^t, o_c^{t+1})$ that corresponds to $\pi_c^{t+1:h-1} = \pi_c^{t:h-1} \Downarrow_{a,o}$ (the sub-tree of $\pi_c^{t:h-1}$ given a_c^t, o_c^{t+1}). That is, by filling out the definition of back-projection, we get

$$\nu(\langle x_c^t, D_c^{t+1} \rangle) = r_a(x_c^t) + \gamma \sum_{o^{t+1}} \left[\sum_{x_c^{t+1}} \bar{O}(o^{t+1}|a^t, x_c^{t+1}) \bar{T}_{I \rightarrow c}(x_c^{t+1}|\langle x_c^t, D_c^{t+1} \rangle, a_c^t) [\Gamma(\pi_c^{t:h-1}, a_c^t, o_c^{t+1})](\langle x_c^{t+1}, D_c^{t+2} \rangle) \right],$$

where D_c^{t+2} is specified as a function of $x_c^t, D_c^{t+1}, a^t, x_c^{t+1}$. Clearly, introducing a maximization can not decrease the value, so

$$\nu(\langle x_c^t, D_c^{t+1} \rangle) \quad \leq \quad r_a(x_c^t) + \gamma \sum_{o^{t+1}} \sum_{x_c^{t+1}} \bar{O}(o^{t+1}|a^t, x_c^{t+1}) \bar{T}_{I \rightarrow c}(x_c^{t+1}|\langle x_c^t, D_c^{t+1} \rangle, a_c^t) \max_{D_c^{t+2}} [\Gamma(\pi_c^{t:h-1}, a_c^t, o_c^{t+1})](\langle x_c^{t+1}, D_c^{t+2} \rangle)$$

$$\{\text{I.H.}\} \quad \leq \quad r_a(x_c^t) + \gamma \sum_{o^{t+1}} \sum_{x_c^{t+1}} \bar{O}(o^{t+1}|a^t, x_c^{t+1}) \bar{T}_{I \rightarrow c}(x_c^{t+1}|\langle x_c^t, D_c^{t+1} \rangle, a_c^t) [\Gamma_{IO}(\pi_c^{t:h-1}, a_c^t, o_c^{t+1})](x_c^{t+1}) \tag{18}$$

where $\Gamma_{IO}(\pi_c^{t:h-1}, a_c^t, o_c^{t+1})$ is the IO vector that corresponds to $\pi_c^{t+1:h-1} = \pi_c^{t:h-1} \Downarrow_{a,o}$.

Now we define the r.h.s. vector:

$$\nu_{IO}(x_c^t) = r_a(x_c^t) + \gamma \sum_{o^{t+1}} \nu_{IO}^{\pi|ao}(x_c^t)$$

$$= r_a(x_c^t) + \gamma \sum_{o^{t+1}} \left[\max_{u_c^{t+1}} \sum_{x_c^{t+1}} O(o_c^{t+1}|a_c, x_c^{t+1}) \Pr(xn_c^{t+1}|x_c^t, a_c, u_c^{t+1}) \Pr(xl_c^{t+1}|x_c^t, a_c) [\Gamma_{IO}(\pi_c^{t:h-1}, a_c^t, o_c^{t+1})](x_c^{t+1}) \right] \tag{19}$$

We need to show that

$$\max_{D_c^{t+1}} \left[r_a(x_c^t) + \gamma \sum_{o^{t+1}} \sum_{x_c^{t+1}} \bar{O}(o^{t+1}|a^t, x_c^{t+1}) \bar{T}_{I \rightarrow c}(x_c^{t+1}|\langle x_c^t, D_c^{t+1} \rangle, a_c^t) [\Gamma_{IO}(\pi_c^{t:h-1}, a_c^t, o_c^{t+1})](x_c^{t+1}) \right]$$

$$\leq r_a(x_c^t) + \gamma \sum_{o^{t+1}} \max_{u_c^{t+1}} \sum_{x_c^{t+1}} O(o_c^{t+1}|a_c, x_c^{t+1}) \Pr(xn_c^{t+1}|x_c^t, a_c, u_c^{t+1}) \Pr(xl_c^{t+1}|x_c^t, a_c) [\Gamma_{IO}(\pi_c^{t:h-1}, a_c^t, o_c^{t+1})](x_c^{t+1})$$

which holds if and only if

$$\max_{D_c^{t+1}} \sum_{o^{t+1}} \sum_{x_c^{t+1}} \bar{O}(o^{t+1}|a^t,x_c^{t+1})\bar{T}_{I\to c}(x_c^{t+1}|\langle x_c^t,D_c^{t+1}\rangle,a_c^t)[\Gamma_{IO}(\pi_c^{t:h-1},a_c^t,o_c^{t+1})](x_c^{t+1})$$

$$\leq \sum_{o^{t+1}} \max_{u_c^{t+1}} \sum_{x_c^{t+1}} O(o_c^{t+1}|a_c,x_c^{t+1})\Pr(xn_c^{t+1}|x_c^t,a_c,u_c^{t+1})\Pr(xl_c^{t+1}|x_c^t,a_c)[\Gamma_{IO}(\pi_c^{t:h-1},a_c^t,o_c^{t+1})](x_c^{t+1}). \quad (20)$$

To show this is the case, we start with the l.h.s.:

$$\max_{D_c^{t+1}} \sum_{o^{t+1}} \sum_{x_c^{t+1}} \bar{O}(o^{t+1}|a^t,x_c^{t+1})\bar{T}_{I\to c}(x_c^{t+1}|\langle x_c^t,D_c^{t+1}\rangle,a_c^t)[\Gamma_{IO}(\pi_c^{t:h-1},a_c^t,o_c^{t+1})](x_c^{t+1})$$

$$= \max_{D_c^{t+1}} \sum_{o^{t+1}} \sum_{x_c^{t+1}} \bar{O}(o^{t+1}|a^t,x_c^{t+1}) \left[\sum_{u_c^{t+1}} \Pr(xn_i^{t+1}|x_c^t,a_c^t,u_c^{t+1})I(u_c^{t+1}|D_c^{t+1})\right] \Pr(xl_c^{t+1}|x_c^t,a_c^t)[\Gamma_{IO}(\pi_c^{t:h-1},a_c^t,o_c^{t+1})](x_c^{t+1})$$

$$= \max_{D_c^{t+1}} \sum_{o^{t+1}} \sum_{u_c^{t+1}} I(u_c^{t+1}|D_c^{t+1}) \sum_{x_c^{t+1}} \bar{O}(o^{t+1}|a^t,x_c^{t+1}) \Pr(xn_i^{t+1}|x_c^t,a_c^t,u_c^{t+1}) \Pr(xl_c^{t+1}|x_c^t,a_c^t)[\Gamma_{IO}(\pi_c^{t:h-1},a_c^t,o_c^{t+1})](x_c^{t+1})$$

$$= \max_{D_c^{t+1}} \sum_{o^{t+1}} \sum_{u_c^{t+1}} I(u_c^{t+1}|D_c^{t+1}) \left[\sum_{x_c^{t+1}} \bar{O}(o^{t+1}|a^t,x_c^{t+1}) \Pr(xn_i^{t+1}|x_c^t,a_c^t,u_c^{t+1}) \Pr(xl_c^{t+1}|x_c^t,a_c^t)[\Gamma_{IO}(\pi_c^{t:h-1},a_c^t,o_c^{t+1})](x_c^{t+1})\right]$$

$\leq$ {max. of a function is greater than its expectation:}

$$\max_{D_c^{t+1}} \sum_{o^{t+1}} \max_{u_c^{t+1}} \left[\sum_{x_c^{t+1}} \bar{O}(o^{t+1}|a^t,x_c^{t+1}) \Pr(xn_i^{t+1}|x_c^t,a_c^t,u_c^{t+1}) \Pr(xl_c^{t+1}|x_c^t,a_c^t)[\Gamma_{IO}(\pi_c^{t:h-1},a_c^t,o_c^{t+1})](x_c^{t+1})\right]$$

$=$ {no dependence on D_c^{t+1} anymore:}

$$\sum_{o^{t+1}} \max_{u_c^{t+1}} \sum_{x_c^{t+1}} \bar{O}(o^{t+1}|a^t,x_c^{t+1}) \Pr(xn_i^{t+1}|x_c^t,a_c^t,u_c^{t+1}) \Pr(xl_c^{t+1}|x_c^t,a_c^t)[\Gamma_{IO}(\pi_c^{t:h-1},a_c^t,o_c^{t+1})](x_c^{t+1})$$

which is the r.h.s. of (20), the inequality the we needed to demonstrate, thereby finishing the proof. $\qquad\square$