Factored Upper Bounds for Multiagent Planning Problems under Uncertainty with Non-Factored Value Functions*

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

Abstract

Nowadays, multiagent planning under uncertainty scales to tens or even hundreds of agents. However, current methods either are restricted to problems with factored value functions, or provide 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, which would additionally allow for meaningful benchmarking of methods of the latter category. To mitigate this problem, this paper introduces a family of *influence-optimistic* upper bounds for factored Dec-POMDPs without factored value functions. We demonstrate how we can achieve firm quality guarantees for problems with hundreds of agents.

1 Introduction

Planning for multiagent systems (MASs) under uncertainty is an important open research problem in artificial intelligence. The decentralized partially observable Markov decision process (Dec-POMDP) is a framework for addressing such problems. Many recent approaches to solving Dec-POMDPs propose to exploit *locality of interaction* [Nair *et al.*, 2005], also referred to as *value factorization* [Kumar *et al.*, 2011]. However, without making very strong assumptions, such as transition and observation independence [Becker *et al.*, 2003], there is no strict locality: in general the actions of any agent may affect the rewards received in a different part of the system. For large MASs, several heuristic approaches have been proposed [Yin and Tambe, 2011; Kumar *et al.*, 2011; Velagapudi *et al.*, 2011; Varakantham *et al.*, 2012; Oliehoek *et al.*, 2013b; Wu *et al.*, 2013; Varakantham *et al.*, 2014], which come without guarantees, however.

In this work, we mitigate this issue by proposing a novel family of techniques that can be used to provide upper bounds on the performance of large *factored* Dec-POMDPs. In addition to 1) quantifying the performance gap of heuristic methods, computing upper bounds is important for other reasons: Knowledge of performance gaps 2) is crucial for researchers to direct their focus to promising areas and 3) sheds light on which problems are easier to approximate than others. 4) Last, but not least, these bounds serve as admissible heuristics for current and future branch&bound search methods.

Computing upper bounds on the achievable value of a planning problem typically involves relaxing the original problem by making some optimistic assumptions. By exploiting the fact that transition and observation dependence leads to a value function that is *additively factored* into a number of small components, researchers have designed techniques for computing upper bounds in so-called Networked Distributed POMDPs (ND-POMDPs) with many agents [Varakantham *et al.*, 2007; Marecki *et al.*, 2008; Dibangoye *et al.*, 2014]. Unfortunately, the assumption of factored value functions narrows down the applicability of such models, 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 from the rest of the problem, 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, a 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 quality guarantees the bounds provide for heuristic methods and, in an extended version [Oliehoek *et al.*, 2015], use our bounds to prune in an A* search. The proposed bounds give meaningful quality guarantees for factored Dec-POMDPs with hundreds of agents. This is a major accomplishment since previous approaches that provide guarantees 1) have required value factorization [Becker *et al.*, 2003; 2004; Varakantham *et al.*, 2007; Dibangoye *et al.*, 2014] or specific interaction topologies [Witwicki, 2011], 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 others' actions.

*An extended version of this paper is available on ArXiv [Oliehoek *et al.*, 2015].

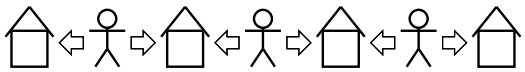

Figure 1: The FIREFIGHTINGGRAPH problem.

2 Background

We focus on Dec-POMDPs where the transition and observation models can be represented compactly as a *two-stage dynamic Bayesian network* (2DBN) [Boutilier *et al.*, 1999]:

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 determine the set of states $\mathcal{S} = \bigotimes_{k=1}^{m} X^k$.
- $T(s'|s, a)$, the transition model specified by a set of conditional probability tables (CPTs), one for each factor.
- $O(o|a, s')$, the observation model: a CPT per agent.
- $\mathcal{R}$ is a set of *local* reward function.
- 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:

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

For instance, Fig. 1 shows the FIREFIGHTINGGRAPH (FFG) problem [Oliehoek *et al.*, 2013b], 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,[1] as illustrated in Fig. 3(left) which shows the 2DBN for a 4-agent instantiation of FFG. The figure shows that the connections are *local* but there is no *transition independence* [Becker *et al.*, 2003] or *value factorization* [Kumar *et al.*, 2011]: 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 or communication networks.

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., a maximum expected sum of rewards: $V(\pi) \triangleq \mathbf{E}[\sum_{t=0}^{h-1} R(s, a, s') \mid b^0, \pi]$. Such an optimal joint policy is denoted π^*.[2]

[1] FFG has rewards of form $R^l(x'_l)$, but we support $R^l(x_l, a_l, x'_l)$ in general.

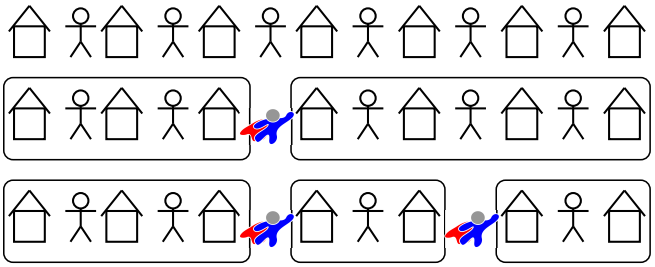

Figure 2: Constructing a global upper bound for 6-agent FFG using influence-optimistic upper bounds for sub-problems.

As explained in the introduction, computing upper bounds is useful for many reasons. For ND-POMDPs, this is tractable for problems with many agents [Varakantham *et al.*, 2007; Dibangoye *et al.*, 2014].[3] However, these methods 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. An upper bound is easily constructed as the sum of local upper bounds: $\hat{V} = \sum_{e \in \mathcal{E}} \hat{V}_e$. The key point is that computing the local upper bounds $\hat{V}_e$ is easy *because* the original problem has a factored value function: each component e can be upper bounded in isolation.

3 Global Upper Bounds

In this paper, we propose a class of *factored* upper bounds, which we call *influence-optimistic upper bounds (IO-UBs)*, for factored Dec-POMDPs that do *not* have a factored value function. This is an important contribution, since currently no methods to compute upper bounds for such problems are available. The overall approach that we take is to divide the problem into *sub-problems* (*SPs*, defined in Section 4), compute overestimations of the achievable value for each of these sub-problems and combine those into a global upper bound.

Combining the local bounds into a global bound is similar to existing methods for computing upper bounds for ND-POMDPs. The basic idea is to apply a non-overlapping decomposition $\mathcal{C}$ (i.e., a partitioning) of the reward functions $\mathcal{R}$ of the original factored Dec-POMDP into SPs $c \in \mathcal{C}$, and to compute a *local* IO-UB $\hat{V}_c^{IO}$ for each (with the methods proposed in Section 4). The *global influence-optimistic upper bound* is then given by: $\hat{V}^{IO} \triangleq \sum_{c \in \mathcal{C}} \hat{V}_c^{IO}$.

We illustrate the construction of a global upper bound $\hat{V}$ for the 6-agent FFG in Fig. 2, which shows the original prob-

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

[3] It would also be possible to compute UBs for TD-POMDPs [Witwicki and Durfee, 2010] with many agents, but here too, existing techniques would require a factored value function, which in turn requires very specific restrictions: interactions must be *directed* and agents can only have a few interaction ancestors.

lem (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 SP c. Effectively, the bounds assume that agent 3 fights fire at both house 3 and house 4 simultaneously (and hence is represented by a superhero figure). Fig. 2 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 on just one side.

To prove correctness of this upper bounding scheme, we define $V_c(\pi)$ as the value realized for R_c, the reward components modeled in SP c, under joint policy π. Given the policies of other agents π_{-c}, the *best-response value* of c is $V_c^{BR}(\pi_{-c}) \triangleq \max_{\pi_c} V_c(\pi_c, \pi_{-c})$. Finally, the *locally-optimal value* for an SP c,

$$V_c^{LO} \triangleq \max_{\pi_{-c}} V_c^{BR}(\pi_{-c}), \tag{1}$$

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 denote the maximizing argument by π_{-c}^{LO}.

Theorem 1. *Let $\mathcal{C}$ be a partitioning of $\mathcal{R}$ into SPs. If $\forall_{c\in\mathcal{C}}\ \hat{V}_c^{IO} \geq V_c^{LO}$ then the global IO-UB is in fact an upper bound to the optimal value $\hat{V}^{IO} \geq V^*$.*

Proof. We have $\hat{V}^{IO} \triangleq \sum_{c\in\mathcal{C}} \hat{V}_c^{IO} \geq \sum_{c\in\mathcal{C}} V_c^{LO} \triangleq \sum_{c\in\mathcal{C}} \max_{\pi_{-c}} V_c^{BR}(\pi_{-c}) = \sum_{c\in\mathcal{C}} \max_{\pi} V_c(\pi_c, \pi_{-c}) \geq \max_{\pi} \sum_{c\in\mathcal{C}} V_c(\pi_c, \pi_{-c}) = V(\pi^*)$. The last equality holds because $\mathcal{C}$ is a partitioning of the reward functions. □

4 Local Upper Bounds

In this section we present our main technical contribution: the machinery to compute a number of *influence-optimistic upper bounds (IO-UBs)* for the value of sub-problems. We first introduce sub-problems formally and then move to techniques for upper bounding their local values, based on theory from influence-based abstraction. While not further explored in this paper, we point out that these techniques can be trivially modified to compute 'pessimistic' influence (i.e., lower) bounds, which could be useful in competitive settings, or for risk-sensitive planning [Marecki and Varakantham, 2010].

Here we introduce two local bounds, based on combining influence-optimism with solution methods for multiagent POMDPs (MPOMDPs) [Messias *et al.*, 2011] and Dec-POMDPs respectively. In the extended version, we also introduce a bound based on multiagent MDPs [Oliehoek *et al.*, 2015]. While this latter bound can be less tight (as demonstrated in the experiments), it is computationally cheaper.

4.1 Sub-Problems

The notion of an SP generalizes the concept of a *local-form model* [Oliehoek *et al.*, 2012] to multiple agents: 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, respectively. 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 $R_c(x_c, a_c, x_c') \triangleq \sum_{l\in\mathcal{R}'} R^l(x_l, a_l, x_l')$.

Figure 3: Left: A 2-agent sub-problem within 4-agent FFG. Right: the corresponding IASP.

However, this is generally not enough to end up with a fully specified (smaller) factored Dec-POMDP. This is illustrated in Fig. 3(left), which shows an SP of FFG involving two agents: 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.[4] We refer to such factors as *non-locally affected factors (NLAFs)* and denote them xn_c^k, where 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 shows that the transitions are not well-defined since the NLAFs depend on the sources of the highlighted *influence links.* We refer to these as *influence sources* $u_c = \langle y_u, a_u \rangle$ (in this case $y_u = \langle X^{i-1}, X^{i+3} \rangle$ and $a_u = \langle a_{i-1}, a_{i+2} \rangle$).

4.2 Influence-Augmented SPs

A local-form model can be transformed to an *influence-augmented local model*, which captures the influence of non-modeld parts of the environment [Oliehoek *et al.*, 2012]. We extend this approach to SPs, leading to *influence-augmented sub-problems (IASPs).* The construction of an IASP consists of two steps: 1) capturing the influence of the non-modeled parts (given π_{-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

[4] We assume that the scopes of the observation CPTs for the included agents are fully contained within the SP, and similarly for the included reward factors. This is not a strong assumption, since situations with arrows pointing to rewards or observations can be modeled by introducing auxiliary state variables. These restrictions generalize previous notions of observation and reward independence [Becker *et al.*, 2003; Nair *et al.*, 2005] (we allow overlap on state factors that can be influenced by the agents themselves). We do *not* assume transition independence, nor do we assume any of the transition-decoupling (i.e., TD-POMDP [Witwicki and Durfee, 2010]) restrictions.

create a model with a transformed transition model $T_{I_{\to c}}$ and no further dependence on the external problem.

Step 1) An *incoming influence point* can be specified as an *incoming influence* $I^t_{\to c}$ for each stage: $I_{\to c} = \left(I^1_{\to c}, \ldots, I^h_{\to c}\right)$. Each such $I^{t+1}_{\to c}$ corresponds to the influence that the SP experiences at stage $t+1$, and thus specifies the conditional probability distribution of the influence sources $u_c = \langle y_u, a_u \rangle$ at stage t. Assuming that the influencing agents use a deterministic policy π_u, $I^{t+1}_{\to c}$ is given by $I(u_c|D_c) = \sum_{\vec{o}_u} \mathbf{1}_{\{a_u = \pi_u(\vec{o}_u)\}} \Pr(y_u, \vec{o}_u | D_c, \vec{b}^0, \pi_{-c})$, with $\mathbf{1}_{\{\cdot\}}$ the Kronecker Delta function, and D_c the *d-separating set* for $I^{t+1}_{\to c}$: the history of a subset of all the modeled variables that d-separates the modeled variables from the non-modeled ones (see [Oliehoek *et al.*, 2012] for details).

Step 2) defines the IASP $\mathcal{M}^{IA}_c = \langle \mathcal{M}_c, I_{\to c} \rangle$ for an SP $\mathcal{M}_c = \langle \mathcal{M}, \mathcal{D}', \mathcal{X}', \mathcal{R}' \rangle$ as a factored Dec-POMDP with the following components. The set of state factors $\bar{\mathcal{X}} = \mathcal{X}' \cup \{D_c\}$ is such that states $\bar{x}_c = \langle x_c, D_c \rangle$ specify a local state of the SP, as well as the d-separating set D_c for the next-stage influences. Only the agents (implying their actions and observations) and rewards from c participate: $\bar{\mathcal{D}} = \mathcal{D}'$ and $\bar{\mathcal{R}} = \mathcal{R}'$. For all OLAFs x^k_c we take the CPTs from the factored Dec-POMDP $\mathcal{M}$, but for all NLAFs we take their *induced CPTs* [Oliehoek *et al.*, 2012], leading to an influence-augmented transition model which is the product of CPTs of OLAFs and NLAFS:

$$\bar{T}_{I_{\to c}}(x'_c | \langle x_c, D_c \rangle, a_c) = \Pr(xl'_c | x_c, a_c) \sum_{u_c = \langle y_u, a_u \rangle} \Pr(xn'_c | x_c, a_c, u_c) I(u_c | D_c). \quad (2)$$

(Note that x_c, a_c, x'_c and D_c together uniquely specify D'_c). The observation model $\bar{O}$ follows directly from O (from $\mathcal{M}$). Fig. 3(right) illustrates the IASP for FFG.

Theorem 2. *$V_c(I_{\to c}(\pi_{-c}))$, the value of an optimal solution of the IASP for influence point $I_{\to c}(\pi_{-c})$, equals the best-response value: $V^{BR}_c(\pi_{-c}) = V_c(I_{\to c}(\pi_{-c}))$.*

Proof. The proof by Oliehoek *et al.* [2012] extends to multi-agent SPs. □

4.3 An MPOMDP-Based Upper Bound

Via (1), it is clear that V^{LO}_c corresponds to the value of the locally optimal influence: $V^{LO}_c = V_c(I_{\to c}(\pi^{LO}_{-c}))$. As such, 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 due to the large number of augmented states $\bar{x}_c = \langle x_c, D_c \rangle$. However, it turns out computing upper bounds to V^{LO}_c can be done more efficiently, without even constructing the IASPs: we can directly use the (under-specified) SPs and modify the 'backup operators' used to compute the optimal value function to make optimistic assumptions about the non-specified influence sources.

The first such upper bounding method we introduce, *influence-optimistic Q-MPOMDP (IO-Q-MPOMDP)*, treats the SP under concern as an (under-specified) multiagent POMDP (MPOMDP) [Messias *et al.*, 2011]. An MPOMDP is partially observable, but assumes that the agents can freely communicate their observations, such that the problem reduces to a centralized one in which a single decision maker (representing the team of agents) takes joint actions, and receives joint observations. The optimal value for an MPOMDP is analogous to that of a POMDP: $Q(b,a) = R(b,a) + \sum_o \Pr(o|b,a) V(b')$, where b' is the joint belief resulting from a Bayes update of b given a and o.

In case that the influence on an SP is fully specified, POMDP techniques can be readily applied to the IASP. However, we want to deal with the case where this influence is not specified. To accomplish this, IO-Q-MPOMDP adds optimistic assumptions on the influences. We propose a formulation that makes use of 'back-projected value vectors': $\nu^{ao}(s) \triangleq \sum_{s'} O(o|a,s') T(s'|s,a) \nu(s')$. (See, e.g., [Spaan, 2012; Shani *et al.*, 2013] for more details.)

The key insight that enables applying influence-optimism to the MPOMDP case is that in this back-projected form we can take the maximum with respect to unspecified influences. We define the *influence-optimistic back-projection* as:

$$\nu^{ao}_{IO}(x_c) \triangleq \max_{u_c} \sum_{x'_c} O(o_c|a_c, x'_c) \Pr(xn'_c|x_c, a_c, u_c) \Pr(xl'_c|x_c, a_c) \nu_{IO}(x'_c). \quad (3)$$

Comparing this equation to (2), it is clear that this equation is optimistic with respect to the influence: it selects the sources u_c in order to select the most beneficial transition probabilities. Since this equation does not depend in any way on the d-separating sets and influence, we can completely avoid generating large IASPs. When combined with an *exact* POMDP solver, such influence-optimistic back-ups will lead to an upper bound $\hat{V}^P_c$ on the locally-optimal value.

Theorem 3. *IO-Q-MPOMDP yields an upper bound on the locally-optimal value: $V^{LO}_c \leq \hat{V}^P_c$.*

Sketch of Proof. The proof shows that for any fixed joint policy π_c, the value vectors $\nu_P \in \mathcal{V}$ and $\nu_{IO} \in \mathcal{V}_{IO}$ (computed under regular- and IO back-projections respectively) satisfy: $\forall_{x_c} \max_{D_c} \nu(\langle x_c, D_c \rangle) \leq \nu_{IO}(x_c)$. This implies that, for any $I_{\to c}$, $V^P_{I_{\to c}}(\bar{b}_{I_{\to c}}) \leq V^{IO}(\bar{b}_{IO})$, provided that the marginal of $\bar{b}_{I_{\to c}}$ coincides with $\bar{b}_{IO}$: $\sum_{D_c} \bar{b}_{I_{\to c}}(\langle x_c, D_c \rangle) = \bar{b}_{IO}(x_c)$, a condition that is satisfied for the initial beliefs. □

4.4 A Dec-POMDP-Based Upper Bound

The previous approach computes upper bounds by, apart from the IO assumption, additionally making optimistic assumptions on communication capabilities. Here we present a method for computing *Dec-POMDP-based upper bounds* that, other than the optimistic assumptions about neighboring SPs, makes no additional assumptions. The approach builds on the recent insight [MacDermed and Isbell, 2013; Dibangoye *et al.*, 2013; Oliehoek *et al.*, 2013a] that a Dec-POMDP can be converted to a special case of POMDP,[5] and

[5] Oliehoek and Amato [2014] give an overview of this reduction.

that therefore we can leverage the IO back-projection (3) to compute an UB, $\hat{V}_c^D$, that we refer to as *IO-Q-Dec-POMDP*.

In particular, we define the *plan-time sub-problem* $\mathcal{M}_c^{PT}$ as the (under-specified) POMDP with states of the form $\check{s} = \langle x_c, \vec{o}_c \rangle$. Each $\check{a}$ corresponds to a local joint decision rule δ_c in the SP (composed of individual decision rules that map observation histories to actions). Rewards are given by $\check{R}(\check{s}, \check{a}) = R_c(x_c, \delta_c(\vec{o}_c))$. There is a single observation $\check{\mathcal{O}} = \{NULL\}$ that is received with probability 1 irrespective of the state and action. The horizon is unmodified: $\check{h} = h$. Finally, the transition model is underspecified, since it depends on the non-specified influence sources: $\check{T}(\check{s}'|\check{s}, \check{a}) = T_{u_c}(x_c'|x_c, \delta_c(\vec{o}_c), u_c) O(o_c|\delta_c(\vec{o}_c), x_c')$.

Since this model is a special case of a POMDP, the theory developed in Section 4.3 applies: we can write down the IO back-projection (3) which in this case translates to[6]

$$\nu^{\delta_c}(x_c, \vec{o}_c) \triangleq \max_{u_c} \sum_{x_c'} \Pr(o_c|\delta_c(\vec{o}_c), x_c') \Pr(xn_c'|x_c, \delta_c(\vec{o}_c), u_c) \Pr(xl_c'|x_c, \delta_c(\vec{o}_c)) \nu(x_c', \vec{o}_c'). \quad (4)$$

Theorem 4. *IO-Q-Dec-POMDP yields an upper bound to the locally-optimal value:* $V_c^{LO} \leq \hat{V}_c^D$.

Proof. Directly from the applying Theorem 3 to $\mathcal{M}_c^{\text{PT}}$. □

4.5 Complexity Analysis

Due to the maximization in (3) and (4), IO back-projections are more costly than regular (non-IO) back-projections. In particular, the complexity of each backup is multiplied by the number of influence source instantiations $|u_c|$. As such, the relative overhead, when compared to solving the SPs as regular (non-IO) MPOMDPs and Dec-POMDPs, is equal for both methods [Oliehoek *et al.*, 2015].

5 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. We compare the influence-optimistic MPOMDP and Dec-POMDP UBs, $\hat{V}_c^P$ and $\hat{V}_c^D$, as well as the IO multiagent MDP (MMDP) bound, $\hat{V}_c^M$, based on the same technique [Oliehoek *et al.*, 2015]. While we focus on the (relative) values found by these heuristics, as the analysis of Section 4.5 indicates that relative timing results follow those of regular (non-IO) MPOMDP and Dec-POMDP methods, we do provide some indicative running times.[7] The extended version [Oliehoek *et al.*, 2015] provides preliminary evidence that our bounds can be used to improved heuristic influence search [Witwicki *et al.*, 2012].

Comparison of Different Bounds. Although the approach described in the paper is general, in the numerical evaluation here we exploit the benchmarks' 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 transition model for the first house (say X^i in Fig. 3) is 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. 3).

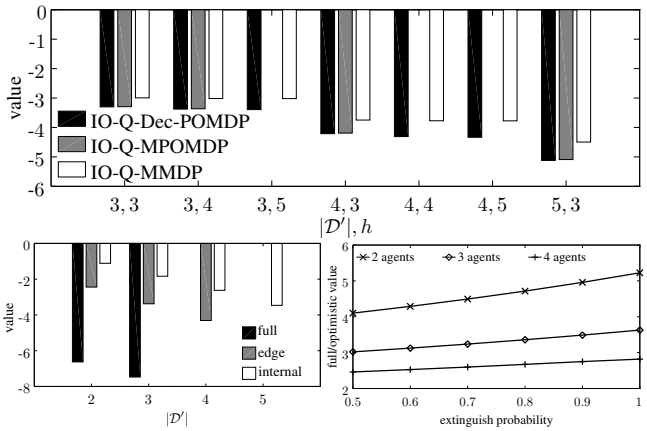

Figure 4: Comparison of different bounds (see text).

Fig. 4(top) 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 $\hat{V}_c^D$ and $\hat{V}_c^P$ is small, but these could become larger for longer horizons. The same analysis for the ALOHA benchmark [Oliehoek *et al.*, 2013b] gave similar results.

We also compare different types of SPs (internal and edge cases, see Fig. 2) encountered in FFG ($h = 4$). In addition, Fig. 4(left) 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.) These results demonstrate a potentially large effect of influence-optimism: 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 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; FFG has a high *influence strength.*

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 orig-

[6] Note that $O(o_i'|a_i, x_i')$ in (3) corresponds to the *NULL* observation in the PT model, but since the observation histories are in the states, $\Pr(o_c|\delta_c(\vec{o}_c), x_c')$ comes out of the *transition* model.

[7] All experiments are run on an Intel Xeon E5-2650L, 32GB system making use of one core only.

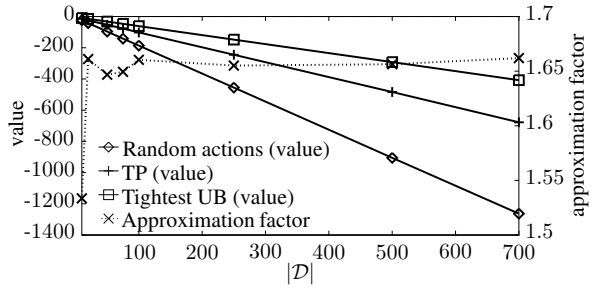

Figure 5: The global IO-Q-Dec-POMDP upper bound on large FFG instances.

inal 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 that there is a clear relation between the fire-extinguish probability when two agents fight fire at a house, and the ratio between the 'full' value (the Dec-POMDP value) and optimistic value.

Bounding the Error of 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 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) [Oliehoek *et al.*, 2013b], one of the methods capable of providing solutions for large factored Dec-POMDPs. Since this heuristic method 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. 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)*: $\max\{\hat{V}^{IO}/V^{TP}, V^{TP}/\hat{V}^{IO}\}$, a number comparable to the approximation factors of *approximation algorithms* [Vazirani, 2001].

Following this methodology, 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 far from the upper bound. In particular, the EAF lies typically between 1.4 and 1.7, thus demonstrating that IO-UBs can provide firm guarantees for solutions of factored Dec-POMDPs with up to 700 agents. Moreover, we see that 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.

$\lvert\mathcal{D}\rvert$	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: ALOHA: Empirical approximation factors for $h = 3$.

Table 1 shows results obtained for ALOHA with up to $|\mathcal{D}| = 250$ agents making use of SPs involving up to $|\mathcal{D}'| = 6$ agents. The numbers clearly illustrate that it is possible to provide very strong guarantees for problems up to 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 Related and Future Work

One way or another, all upper bounds in the literature make some optimistic assumption, but *influence-optimistic* UBs are novel. While the idea of being optimistic *with respect to influences* has been considered by Kumar and Zilberstein [2009a], they do not provide an upper bound the global value. Instead they employ optimistic assumptions on transitions in an ND-POMDP to derive an MMDP-based policy which is used to sample belief points. Kumar and Zilberstein [2009b] present the only previous method that delivers scalability with respect to the number of agents without assuming value factorization by exploiting submodular function maximization for a specific class of sensor network problems.

Upper bounds for ND-POMDPs (e.g., [Varakantham *et al.*, 2007; Dibangoye *et al.*, 2014]) resemble our global IO upper bound. 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. A promising idea is to employ our factored upper bounds in combination with the heuristic search methods by Dibangoye *et al.* 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.

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 provide preliminary evidence in the extended version of this paper [Oliehoek *et al.*, 2015]), and A* methods for a large class of factored Dec-POMDPs (as mentioned above).

7 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 [Oliehoek *et al.*, 2012], 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 (NEXP-complete [Rabinovich *et al.*, 2003]) 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.

Acknowledgments

F.O. is supported by NWO Innovational Research Incentives Scheme Veni #639.021.336.

References

[Becker *et al.*, 2003] R. Becker, S. Zilberstein, V. Lesser, and C. V. Goldman. Transition-independent decentralized Markov decision processes. In *AAMAS*, 2003.

[Becker *et al.*, 2004] R. Becker, S. Zilberstein, and V. Lesser. Decentralized Markov decision processes with event-driven interactions. In *AAMAS*, 2004.

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

[Dibangoye *et al.*, 2013] J. S. Dibangoye, C. Amato, O. Buffet, and F. Charpillet. Optimally solving Dec-POMDPs as continuous-state MDPs. In *IJCAI*, 2013.

[Dibangoye *et al.*, 2014] J. S. Dibangoye, C. Amato, O. Buffet, and F. Charpillet. Exploiting separability in multiagent planning with continuous-state MDPs. In *AAMAS*, 2014.

[Kumar and Zilberstein, 2009a] A. Kumar and S. Zilberstein. Constraint-based dynamic programming for decentralized POMDPs with structured interactions. In *AAMAS*, 2009.

[Kumar and Zilberstein, 2009b] A. Kumar and S. Zilberstein. Event-detecting multi-agent MDPs: Complexity and constant-factor approximations. In *IJCAI*, 2009.

[Kumar *et al.*, 2011] A. Kumar, S. Zilberstein, and M. Toussaint. Scalable multiagent planning using probabilistic inference. In *IJCAI*, 2011.

[MacDermed and Isbell, 2013] L. C. MacDermed and C. Isbell. Point based value iteration with optimal belief compression for Dec-POMDPs. In *NIPS 26*. 2013.

[Marecki and Varakantham, 2010] J. Marecki and P. Varakantham. Risk-sensitive planning in partially observable environments. In *AAMAS*, 2010.

[Marecki *et al.*, 2008] J. Marecki, T. Gupta, P. Varakantham, M. Tambe, and M. Yokoo. Not all agents are equal: scaling up distributed POMDPs for agent networks. In *AAMAS*, 2008.

[Messias *et al.*, 2011] J. V. Messias, M. T. J. Spaan, and P. U. Lima. Efficient offline communication policies for factored multiagent POMDPs. In *NIPS 24*. 2011.

[Nair *et al.*, 2005] R. Nair, P. Varakantham, M. Tambe, and M. Yokoo. Networked distributed POMDPs: A synthesis of distributed constraint optimization and POMDPs. In *AAAI*, 2005.

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

[Oliehoek *et al.*, 2012] F. A. Oliehoek, S. Witwicki, and L. P. Kaelbling. Influence-based abstraction for multiagent systems. In *AAAI*, 2012.

[Oliehoek *et al.*, 2013a] 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 Artificial Intelligence Research*, 46:449–509, 2013.

[Oliehoek *et al.*, 2013b] F. A. Oliehoek, S. Whiteson, and M. T. J. Spaan. Approximate solutions for factored Dec-POMDPs with many agents. In *AAMAS*, 2013.

[Oliehoek *et al.*, 2015] F. A. Oliehoek, M. T. J. Spaan, and S. Witwicki. Influence-optimistic local values for multiagent planning — extended version. *ArXiv e-prints*, arXiv:1502.05443, February 2015.

[Rabinovich *et al.*, 2003] Z. Rabinovich, C. V. Goldman, and J. S. Rosenschein. The complexity of multiagent systems: the price of silence. In *AAMAS*, 2003.

[Shani *et al.*, 2013] 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.

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

[Varakantham *et al.*, 2007] 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 *AAMAS*, 2007.

[Varakantham *et al.*, 2012] P. Varakantham, S. Cheng, G. J. Gordon, and A. Ahmed. Decision support for agent populations in uncertain and congested environments. In *AAAI*, 2012.

[Varakantham *et al.*, 2014] P. Varakantham, Y. Adulyasak, and P. Jaillet. Decentralized stochastic planning with anonymity in interactions. In *AAAI*, 2014.

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

[Velagapudi *et al.*, 2011] P. Velagapudi, P. Varakantham, P. Scerri, and K. Sycara. Distributed model shaping for scaling to decentralized POMDPs with hundreds of agents. In *AAMAS*, 2011.

[Witwicki and Durfee, 2010] S. J. Witwicki and E. H. Durfee. Influence-based policy abstraction for weakly-coupled Dec-POMDPs. In *ICAPS*, 2010.

[Witwicki *et al.*, 2012] S. Witwicki, F. A. Oliehoek, and L. P. Kaelbling. Heuristic search of multiagent influence space. In *AAMAS*, 2012.

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

[Wu *et al.*, 2013] F. Wu, S. Zilberstein, and N. R. Jennings. Monte-Carlo expectation maximization for decentralized POMDPs. In *IJCAI*, 2013.

[Yin and Tambe, 2011] Z. Yin and M. Tambe. Continuous time planning for multiagent teams with temporal constraints. In *IJCAI*, 2011.