An Extension of Bayesian Game Approximation to Partially Observable Stochastic Games with Competition and Cooperation.

Laura E. Ray

Dartmouth College

89 PUBLICATIONS **1,174** CITATIONS

SEE PROFILE

Jerald D. Kralik

Korea Advanced Institute of Science and Technology

84 PUBLICATIONS **2,761** CITATIONS

SEE PROFILE

An Extension of Bayesian Game Approximation to Partially Observable Stochastic Games with Competition and Cooperation

Dongqing Shi, Michael Z. Sauter, Xueqing Sun, Laura E. Ray and Jerald D. Kralik

Abstract— **The real world is full of ambiguity. Either an individual agent or a team of agents must make decisions in the presence of various uncertainties. This paper examines how each agent in a decentralized team makes decisions when it is subject to environmental and teammate uncertainties, where the interacting environment is only partially observable and the teammate decisions are not known. The problem is modeled as a partially observable stochastic game (POSG). Due to the NEXP-completeness of finding an optimal solution to POSGs, approximate approaches have been developed for practical applications. Bayesian game approximation has also been applied to solve POSGs. However, current Bayesian game approximation algorithms have only been applied to POSGs with common payoffs. Hence, only cooperation among players has been addressed. In reality, many tasks require both cooperation and competition. In order to fill this gap, this paper extends the current Bayesian game approximation to POSGs with distinct payoff mechanisms among players. In addition, by simulating social problem-solving behaviors of chimpanzees, we also introduce a biologically inspired method to improve the efficiency of the approximation. The simulation results have shown both cooperative and competitive behaviors for a multi-agent team.**

I. Introduction

An environment with elements of uncertainty can be modeled using a partially observable Markov decision process (POMDP). In single-agent cases, a POMDP generalizes the fully observable continuous-space Markov decision process (MDP) by introducing a belief state, which is a probability distribution over the possible states of the environment, based on observations of the environment [6]. The optimal solution of such POMDPs can be solved using dynamic programming or reinforcement learning [14]. In multi-agent cases, however, state transitions are determined not only by the individual agent's actions but by the actions of all the other agents (i.e., the joint action space), which are also uncertain. Thus, a generalized belief state has been introduced, which is based not only on the underlying state but also the policies of the other agents, and corresponding dynamic programming has been applied to this generalized model [7]. It has been shown, however, that an optimal solution to multi-agent planning with uncertainty is NEXP-complete [1], indicating that any optimal algorithm will be computationally intractable with an increase in the number of agents and/or the control horizon.

Researchers have also developed other methods to approximate optimal solutions for multi-agent task planning with uncertainty. Some [9] [11] use policy search algorithms to find local optimality for decentralized POMDPs (DEC-POMDPs). Others [4] [10] use game theory to solve cooperative, partially observable stochastic games (POSGs) with common payoffs. In particular, the common-payoff POSG is solved by transforming it into a series of Bayesian games [4], in which the locally optimal solution, the Bayesian Nash equilibrium, is found through alternating maximization. Alternating maximization is a method whereby each agent takes a turn finding its optimal policy while the other agents' policies are fixed. However, most work is limited to common-payoff POSGs and its equivalent, DEC-POMDPs. As a result, competitive behaviors are not encouraged, even though competition does exist in many practical problems. As an example, in an auction task, there can only be one winner for each bidding item; the loser gets nothing.

This paper considers a more general POSG with distinct payoffs among players. By developing a new heuristic Q_{MMDP}, the Q values of the corresponding fully observable multi-agent Markov decision process, we extend the existing Bayesian game approximation to POSGs with distinct payoffs. This extension can solve POSGs involving both cooperation and competition. For our experiments, we have chosen to simulate chimpanzee foraging because it is a highly complex multi-agent problem that requires both cooperation and competition. By studying this system, we hope to obtain insights into how such difficult problems can be solved. In this paper, we focus on forming teams based on social hierarchy and preferred coalitional partners, which we will call "friendship". These social constructs not only encourage instances of both cooperation and competition, but also help to simplify the original complex problem by decomposing it into smaller problems. Each smaller problem is then solved using the extended Bayesian game approximation. Note that the direct use of POSGs on the overall simulated chimpanzee society is computationally intractable due to the large group size and complexity of the environment, but yet becomes manageable when broken down into these smaller game units.

The remainder of the paper is organized as follows. In section II, we first review social foraging behavior of chimpanzees. In section III, we briefly introduce the definition of POSGs and Bayesian games. In section IV, we describe the POSG and an extended Bayesian game approximation

This work was supported by the Office of Naval Research under Grant No. N00014-08-1-0693.

Dongqing Shi, Michael Z. Sauter and Jerald D. Kralik are with the Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH 03755. {Dongqing.Shi, Michael.Sauter, Jerald.D.Kralik}@Dartmouth.edu

Xueqing Sun and Laura E. Ray are with the Thayer School of Engineering, Dartmouth College, Hanover, NH 03755. {xueqing.sun, laura.e.ray}@Dartmouth.edu

solution in detail. In section V, we present simulation results. Finally, we conclude the paper and discuss future directions.

II. Chimpanzee Behavior

In this section we provide an overview of chimpanzee foraging behavior (see [12] [13] for a more detailed description of mammalian cooperative behavior).

Currently, artificial multi-agent systems underperform social mammals in a number of ways, especially with respect to processing speed, adaptability, and scalability. Nature, thus, can provide insights into how to solve complex problems where multiple individuals coexist and flourish. In order to be robust enough to handle real-world applications, multi-agent systems need to be able to make both group and individual decisions; and indeed, social mammals have evolved to maintain a communal structure that can be both cooperative and competitive at the same time. In addition, group formation needs to be dynamic, with the ability to form smaller or larger groups depending on the environment. The best examples of this are "fission-fusion" societies, in which coalitions form or break apart effortlessly, all dependant on what the environment dictates [8]. In our research, we are focusing on a quintessential fission-fusion society, that of the chimpanzee, with the goal of obtaining insights into how multi-agent problems are successfully managed; and more specifically, here, we are focusing on chimpanzee foraging.

A typical foraging expedition starts with chimpanzees that must make decisions about where to forage and what the size and make up of the group should be, based on the current (assumed) state of the environment and social factors described below. The foraging parties then go out in search of food, with the food locations, quantity and quality being approximately known, but uncertain. Finally, the chimpanzees may break up into even smaller groups based on a number of factors, including whether the group size is supported by a particular patch they encounter, and whether individuals will actually obtain food, especially if they are lower in the dominance hierarchy [2]. This paper draws especially upon the factors that cause the fission of the larger groups into smaller ones.

In order for a multi-agent system to function within the time and computational constraints necessary in the real world, heuristics appear to be needed to reduce the state space. Chimpanzees do this inherently, with social factors that guide their behavior based on a dominance hierarchy, group size, and positive affiliations between certain members of the group. Dominance hierarchy is a prominent factor in the lives of chimpanzees, playing a part in both food and mate selection; every member is aware of not only their place in the hierarchy, but also the rank of the others as well. Grouping is inherently preferred, with group size being determined largely by the safety inherent in numbers, with larger generally being better. However, given the patchiness of the ecology (e.g. fruiting trees), the largest group sizes cannot normally be supported, and therefore the chimpanzees must break up into smaller parties. The third and final social variable considered here helps determine the make up of the parties: strong affiliations between certain members of the group, which we will call "friendship" for simplicity, although this is typically friends or relatives. Unique among animals, only chimpanzees and humans have a male-based society, in which males remain in the community for life, whereas females move to other communities at adolescence (this is less clear in human society today, but is seen throughout history and in many indigenous societies). Thus, male chimpanzees are often related, and closer relatives often tend to be preferred coalition partners. Positive affiliations, or friendships, also seem to form from common goals that arise when two or more chimpanzees share positively rewarding experiences over time (such as working together to fend off others for food). Nonetheless, even though these positive affiliations and grouping preferences exist, the dominance hierarchy remains fairly strict and aggression is relatively high. Thus, in combination, these social factors lead to both cooperation and competition within the society.

The research reported here is inspired by these innate social characteristics that chimpanzees and other social mammals possess, which appear to allow them to solve otherwise intractable multi-agent problems. In the next section, we present the analytical framework used to model chimpanzee foraging behavior.

III. Technical Background

In this section, we briefly introduce the framework of partially observable stochastic games and Bayesian game theory.

A. Partially Observable Stochastic Games

Stochastic games provide a generalization of single-agent systems to multi-agent ones, in which any individual agent's decision must also take the potential actions of others into account; and POSGs are the extension of stochastic games to handle uncertainty. A POSG can be defined as a tuple $(n, S, A, T, R, \Omega, O)$, where n is the number of agents; S is the set of environmental states; A is the set of all actions of all agents, and A_i is the set of all actions for agent i, i.e. $A = \times_{i=1}^{n} A_i$, specifically, a joint action $a =< a_1, \ldots, a_n >$; T is the transition function, $S \times A \times S \rightarrow [0, 1]$; R is the payoff function, $S \times A \rightarrow \mathbb{R}$, and generally $R_i \neq R_j$; Ω is the joint observations of all the agents, and Ω_i is the set of observations of an individual agent i, i.e. $\Omega = \times_{i=1}^{n} \Omega_i$, specifically, a joint observation $\omega =< \omega_1, \ldots, \omega_n >$; and O is the observation function, $S \times A \times \Omega \rightarrow [0, 1]$, representing the likelihood of obtaining the current observation, given the previous state and set of agent actions taken. To solve the POSG is to find an optimal joint policy σ, a Nash equilibrium, that maximizes the payoffs for each agent. The policy of agent i is specified by σ_i, and hence, $\sigma = \times_{i=1}^{n} \sigma_i$. Note that DEC-POMDP is a special POSG with common payoffs.

Theoretically, the extensive form [5] can be used to solve the POSG. However, the game tree expands very rapidly with the increase in the number of agents and/or control horizon; and thus the POSG could be computationally intractable even for small two-agent problems. Researchers [4], however, have been able to approximately solve the POSG by decomposing it into a series of relatively simple Bayesian games. This paper extends this idea to solve larger, more general POSGs by integrating the intelligent social mammalian strategies discussed in Section II with chimpanzees.

B. Bayesian Games

Bayesian games [5] are one-stage games with incomplete information. Players have private information relevant to their decision-making; the private information in general is called "type", which can include, for example, one's personal observations: for instance, in chimpanzee foraging, a particular chimpanzee's type could be its observation of which food patch has the highest quality fruit. A Bayesian game can be defined as a tuple (n, A, Θ, P, μ), where n is the number of players; A is the set of all actions of all the agents; Θ is the joint type, and Θ_i is the set of all possible types for player i, and θ_i is a realization of Θ_i, i.e. $\Theta = \times_{i=1}^{n} \Theta_i,$ and $\theta_i \in \Theta_i$; P is the probability distribution over the joint types, $\Theta \to [0, 1]$, given that any agent i is uncertain about the other agents' types, since type is considered private information; μ is the payoffs of players, and μ_i is the payoff for the player i, $\mu = \bigcup_{i=1}^{n} \mu_i$. Let θ_{-i} denote the players' types except for player i and σ_{-i} denote the players' policies expect for player i. $\sigma_i^{\Theta_i}$ is the possible policy set of player i. The Nash equilibrium of Bayesian games maximizes the expected payoff for each player i based on its type θ_i:

$$\sigma_i^*(\theta_i) = \underset{\sigma_i \in \sigma_i^{\Theta_i}}{\operatorname{argmax}} \sum_{\theta_{-i} \in \Theta_{-i}} p(\theta_{-i}|\theta_i)\mu_i(\sigma_i, \sigma_{-i}, (\theta_i, \theta_{-i})). \tag{1}$$

This equilibrium is often called a Bayesian equilibrium or a Bayesian Nash equilibrium (see, e.g., [5]).

IV. A Solution to POSGs with Distinct Payoffs

This section proposes an algorithm for solving general POSGs. In our method, team players (i.e. foraging parties) are first grouped into smaller coalitions based on a biologically inspired algorithm. Then each subgroup plays a relatively small POSG that can be approximately solved by Bayesian game decomposition with the introduction of a new heuristic. The following subsections present the details for each step of the procedure.

A. Grouping Algorithms

In general, any grouping decision process, such as voting, could be used [15]; however, most of the grouping algorithms require substantial communication, which may not be robust in practice. Little evidence of much communication is actually found in social animals. Instead, some sort of "leadership" often plays an important role in group decision-making. In this paper, we explore four main influences on grouping in chimpanzees, one based on general motivation, the other three on the key social factors of leadership, group size preference, and preferred coalitional partners, which we generally call "friendship". First, we exploit the fact that not all chimpanzees wish to forage at any given time. Following a simple homeostasis model, each individual in our simulations has different types of motivations (including a drive to forage, patrol, hunt, etc.), and thus they may be motivated to forage at different times. In a small, hypothetical community of ten, for instance, perhaps only seven are motivated to forage at a particular time. Second, because chimpanzees naturally prefer to group, group size will be as large as the ecology of their territory allows. In the simulations reported here, we have simplified the grouping variable, by having each foraging party (described further below) form a group of at most four individuals. One of our next developments will be to make group size a function of ecological constraints (see section VI).

The final two factors are leadership, i.e. dominance, and friendship. These two factors strongly influence grouping and are modeled in the following way. Let D denote the set of dominance rankings, and D_i is the dominance ranking knowledge of the other chimpanzees known by chimpanzee i, i.e.

$$\begin{aligned} D &= \{D_1, \ldots, D_i, \ldots, D_m\}, \\ D_i &= \{d_{i,1}, \ldots, d_{i,j}, \ldots, d_{i,i-1}, d_{i,i+1}, \ldots, d_{i,m}\}, \end{aligned}$$

where, $d_{i,j}$ is the dominance ranking of chimpanzee j known by chimpanzee i and m is the number of chimpanzees. Let F denote the set of friends, and F_i is the level of friendship of chimpanzee i to the other chimpanzees. Similarly,

$$\begin{aligned} F &= \{F_1, \ldots, F_i, \ldots, F_m\}, \\ F_i &= \{f_{i,1}, \ldots, f_{i,j}, \ldots, f_{i,i-1}, f_{i,i+1}, \ldots, f_{i,m}\}. \end{aligned}$$

We assume dominance ranking and friendship information are known in prior. In reality, dominance rankings and friendships change over time, and this knowledge would also be required to update over time. This will be another future development. The current grouping algorithm, then, works in the following way. For all chimpanzees motivated to forage, the most dominant chimpanzee l is known as the leader. By referring to its friendship table F_l and its observation of the environment (described below), the leader picks the best team. Currently, the leader chooses the top three individuals in its friendship table. Once this group vacates, the most dominant chimpanzee of the remaining individuals becomes the new current leader, l, and it then selects its group. This procedure repeats until all teams are formed. Note that this grouping process is a natural form of cooperation, although in future developments, one could imagine the more subordinate individuals vying for friendship with the more dominant ones.

B. Extended Bayesian Game Approximation

The main advantage of the biologically inspired grouping algorithms is that they significantly reduce the size of the problem, which in the present case is determining which of several foraging sites each individual in the foraging party should visit. Here we focus on solving the POSG through solving a sequence of Bayesian games; in section V we will apply it more specifically to chimpanzee foraging. Assume there are n players on a team (i.e. in a foraging party). Equation 1 shows that there are two components needed to determine the Bayesian equilibrium: the conditional probability $p(\theta_{-i}|\theta_i)$ and payoff $\mu_i(\sigma_i, \sigma_{-i}, (\theta_i, \theta_{-i}))$.

Using conditional probability theory, $p(\theta_{-i}|\theta_i)$ can be computed:

$$p(\theta_{-i}|\theta_i) = \frac{p(\theta_{-i}, \theta_i)}{p(\theta_i)}. \tag{2}$$

where $p(\theta_{-i}, \theta_i)$ is the probability distribution over the joint type, and $p(\theta_i)$ is the marginal probability that can be computed thus:

$$p(\theta_i) = \sum_{\theta_{-i} \in \Theta_{-i}} p(\theta_{-i}, \theta_i). \tag{3}$$

Assume E is the event space, with $B_1, B_2, \ldots, B_m$ as a set of known disjointed events, such that $E = \bigcup_j^m B_j$. For example, B_1 could be a particular configuration of food patches on one particular iteration or day, B_2 could be a different configuration, etc. Then, the joint probability of all agents' types is:

$$p(\theta_{-i}, \theta_i) = \sum_{j=1}^{m} p(\theta_{-i}, \theta_i | B_j) p(B_j), \tag{4}$$

where $p(B_j)$ is given in prior, and

$$\begin{aligned} p(\theta_{-i}, \theta_i | B_j) &= p(\theta_{-i}, \theta_i | B_j) \\ &= p(\theta_1, \ldots, \theta_n | B_j) \\ &= p(\theta_1 | B_j) \ldots p(\theta_n | B_j), \end{aligned} \tag{5}$$

where $p(\theta_i|B_j)$ can be obtained from the observation function O. Then equation 2 becomes the following:

$$p(\theta_{-i}|\theta_i) = \frac{\sum_{j=1}^{m} p(\theta_1|B_j) \ldots p(\theta_n|B_j)}{\sum_{\theta_{-i} \in \Theta_{-i}} p(\theta_{-i}, \theta_i)}. \tag{6}$$

And thus the conditional probability $p(\theta_{-i}|\theta_i)$ is fully defined.

As pointed out by [4] [10], the optimal solution to POSGs requires computing optimal payoffs of the sequence of Bayeisan games; and computing these optimal payoffs is impractical without knowing an optimal joint policy in advance. Hence, heuristics such as Q_{MDP} and Q_{POMDP} [10] are used for approximation. However, these heuristics can only be applied to common-payoff POSGs because their underlying MDP is a single agent MDP. New heuristics are needed for POSGs with distinct payoffs. We introduce such a new heuristic value Q_{MMDP}, which are the Q values of the corresponding fully observable multi-agent MDP. Since multiple agents coordinate in a fully observable game, an individual payoff function can be associated with each agent. Let $Q_i(s, a)$ denote the expected future payoff of state s and action a. The values can be calculated using standard dynamic programming techniques for multi-agent MDPs or proper reinforcement learning approaches. We obtain the $Q_i(s, a)$ values through reinforcement learning using WoLF-PHC [3] [13], a policy hill-climbing learning algorithm that is able to find optimal solutions for fully observable multi-agent tasks, using a win or learn fast (WoLF) heuristic. Then the suitable payoffs of Bayesian games can be approximated by

$$\mu_i(\sigma_i, \sigma_{-i}, (\theta_i, \theta_{-i})) \approx \sum_{s^t \in S} Q_i^t(s, a) b^t(s|\theta_i, \theta_{-i}), \tag{7}$$

where $b^t(s|\theta_i, \theta_{-i})$ is the joint belief over states given the joint observation and can be computed as follows:

$$\begin{aligned} b^t(s|\theta_i, \theta_{-i}) &= p(s^t|\theta^t, b^0) \\ &= \frac{p(s^t, \theta^t|b^0)}{p(\theta^t|b^0)}, \end{aligned} \tag{8}$$

where, $\theta^t = (\theta_i^t, \theta_{-i}^t)$; and

$$p(s^t, \theta^t|b^0) = \sum_{s^{t-1} \in S} O \cdot T \cdot p(s^{t-1}, \theta^{t-1}|b^0); \tag{9}$$

where O is the observation function, $p(\omega^t|a^{t-1}, s^t)$; and T is the state transition function, $p(s^t|s^{t-1}, a^{t-1})$. For stage 0, $p(s^{t-1}, \theta^{t-1}|b^0) = p(s^0, \theta^0|b^0) = b^0(s^0)$ and b^0 is the initial joint belief. Finally, by following equation 4,

$$p(\theta^t|b^0) = \sum_{j=1}^{m} p(\theta_{-i}, \theta_i | B_j, b^0) p(B_j|b^0). \tag{10}$$

By specifying the two components of Equation 1 as shown in Equations 6 and 7, POSGs can be computed approximately by alternating finding the maximum payoff of each agent and searching from the most dominant agent to the least dominant agent.

V. SIMULATION

This section first describes the foraging domain, and then presents the simulation results.

A. Foraging Domain

In the "Foraging World", seven chimpanzee-based agents that are motivated to forage coordinate (both cooperatively and competitively) to maximize their individual payoffs, which we initially introduced in [13]. The task is completed when either every agent has obtained food, or there is no food left in the world. The world, as shown in Fig. 1, is made up of six "Food Patches" and an initial "Home" state. Each "Food Patch" (FP) is characterized by four factors: number of food items in a given patch, called patch size (PS), food quality (FQ), risk (RK) associated with the patch (determined by distance to the border of the world), and effort (ET), based on the distance away from the patch. For convenience, FP → (PS, FQ, RK, ET). The food patch values are in the following order:

$$FP_5 > FP_6 > FP_4 > FP_3 > FP_2 > FP_1 \tag{11}$$

The payoff function for each individual depends on the four food patch characteristics, as well as on relative dominance ranking and a safety factor, which is determined by

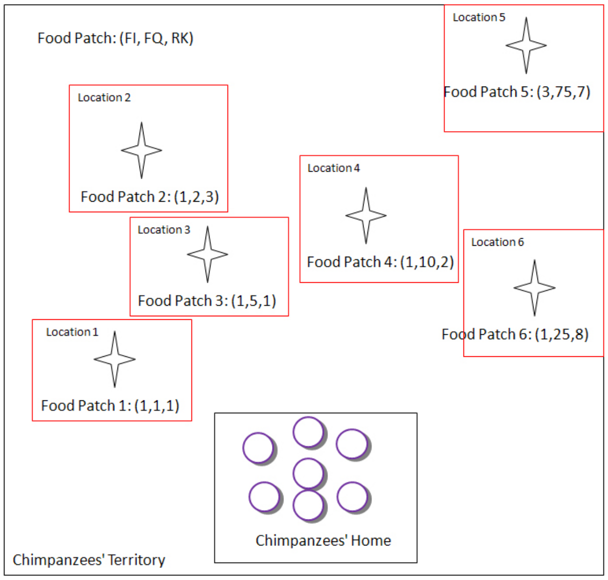

Fig. 1. The Foraging World. Seven chimpanzees are represented by circles, six food patches by stars. The three numbers associated with each patch represent patch size, food quality, and risk. See text for details.

TABLE I

FOOD PATCH (FP_i) OCCURRENCE PROBABILITY P_f AT EACH POSSIBLE LOCATION (L_j).

	L_1	L_2	L_3	L_4	L_5	L_6
FP_1	0.9	0.02	0.02	0.02	0.02	0.02
FP_2	0.02	0.9	0.02	0.02	0.02	0.02
FP_3	0.02	0.02	0.9	0.02	0.02	0.02
FP_4	0.02	0.02	0.02	0.9	0.02	0.02
FP_5	0.02	0.02	0.02	0.02	0.9	0.02
FP_6	0.02	0.02	0.02	0.02	0.02	0.9

TABLE II

OBSERVATION PROBABILITY O OF A VALID SIGNAL (VS) OR NO SIGNAL (NS) FROM EACH FOOD PATCH (FP_i).

	FP_1	FP_2	FP_3	FP_4	FP_5	FP_6
VS	0.2	0.21	0.43	0.45	0.97	0.73
NS	0.8	0.79	0.57	0.55	0.03	0.27

the foraging group size. The payoff is different for agents at the same food patch: the most dominant one is guaranteed the best quality food, whereas more subordinate ones may not obtain anything if the quantity is limited. In detail, the overall payoff is made up of the following five components:

- *Reward_FQ:* Food reward quality (positive) - This value increases based on the quality of the food type. This value is affected by relative rank and patch size.
- *Reward_RK:* Risk at the given food patch (negative) - regardless of whether the agent obtains food or not.
- *Reward_ET:* Effort for traveling to the food patch (negative) - regardless of whether the agent obtains food or not, given that traveling was required.
- *Reward_Hungry:* Satisfaction reward (positive) - if the agent actually obtains food.
- *Reward_Group:* Group size reward (positive) - if the agent forages in a group.

B. POSG Foraging Model

In the Foraging World, the agents know the locations of food patches in general, however, the characteristics of the food patches are uncertain. The six food patches occupy six locations L_i based on the occurrence probability P_f. Only one food patch will appear in a given location and six food patches must appear simultaneously in different locations. P_f is given in Table I, which can also be used to obtain the state transition probability T. In general, each food patch can emit a signal and the probability of emitting a valid signal is proportional to its four patch characteristics: PS, FQ, RK, and ET. Assume the valid signal can always be observed by agents if exists. The observation probability O (signal emission probability) is specified in Table II. Table II reads that six food patches are emitting signals simultaneously: the valid signal (VS) of FP_1 can be observed 20% of the time; the valid signal (VS) of FP_2 can be observed 21% of the time, etc. The "best" food patch FP_5, therefore, has the highest likelihood of being observed. In addition, each agent knows its own type, i.e. what it observes, but does not know the types and actions of the others. Each agent, however, does know the probability distribution over the joint type that can be computed through the occurrence probability P_f and the observation probability O using equations 4 and 5:

$$\begin{aligned} p(\theta_{-i}, \theta_i) &= \sum_{j=1}^{m} p(\theta_{-i}, \theta_i | B_j) p(B_j) \\ &= \sum_{j=1}^{m} p(\theta_1 | B_j) \ldots p(\theta_n | B_j) p(B_j) \\ &= \sum_{j=1}^{m} p(B_j) \Pi_{i=1}^{n} \Pi_{l=1}^{6} O_{i,jl}, \qquad (12) \end{aligned}$$

where $O_{i,jl}$ is the probability over a typical observation i for each food patch l, given a prior event B_j.

To provide a sense of the complexity of the problem space as a full POSG, a quick calculation can be made. For example, the observation instance of agent 1 is $\theta_1 = [L_1 = \text{VS}, L_2 = \text{VS}, L_3 = \text{VS}, L_4 = \text{VS}, L_5 = \text{VS}, L_6 = \text{VS}]$, meaning the agent 1 received a valid signal from location L_1, a valid signal from location L_2, etc. Its observation space Θ_1 has a size of $|\Theta_1|$=2^6=64. Given m agents in the game, the joint observation space has a size of 64^m($\approx$ 4.4 trillion when m=7). It is clear that the problem quickly becomes computationally intractable due to the huge number of possible observations. Thus, it needs to be decomposed into smaller games, which we describe in the next section.

C. Experimental Results

In the experiments, the dominance ranking of the agents is aligned with the agents' indices: the agent with the highest dominance ranking has the highest index. We assume that seven of the chimpanzees in the overall community are motivated to forage. Using the grouping algorithm described in section IV-A, the most dominant agent, Ag 7, chooses the food patch emitting the strongest signal, along with its two neighboring food patches. As illustrated in Fig. 2, Ag 7 chooses L_5 together with L_4 and L_6, and forms its group by selecting its three preferred coalition partners. After the first group heads out, the second group is formed, in which the chimpanzee with the highest remaining rank, in this case Ag 6, selects its preferred foraging partners that remain (Fig. 2). Because the second group has seen the first group head toward L_5, this area will then be avoided. More specifically, in the simulations, the second group will not choose L_5 nor either of the two closest food locations, L_4 or L_6. Ag 6 selects the remaining food patches, and the group then heads toward those patches.

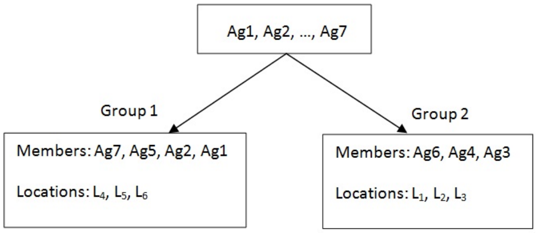

Fig. 2. Two groups determined in turn by the most dominant agent available to join a team. Ag denotes agent. Ag 7 prefers Ag 5, Ag 2, and Ag 1 as its coalition partners, i.e. friends, and they form Group 1. Once the first group is formed, Ag 6 then becomes the most dominant available agent, which then chooses its preferred friends for Group 2.

For every group, once they reach the general area, each individual agent in the group then chooses to forage in either of the three closest food patches, based on the decision algorithm detailed in sections 3 and 4 above, i.e. Equation 1 as approximated by Equations 6 and 7. For Equation 7, Table III shows an example of $Q_i(s, a)$ values obtained through WoLF-PHC multi-agent reinforcement learning when the estimated state $s = [L_4 = FP_4, L_5 = FP_5, L_6 = FP_6]$. Of course, the belief to the state s depends on the joint observation.

TABLE III

THE NORMALIZED $Q_i(s, a)$ VALUES: THE VALUES ARE COLLECTED AFTER THE WOLF-PHC CONVERGES AND THEN NORMALIZED BY THE LARGEST Q VALUES. G_1 AND G_2 REPRESENTS GROUP 1 AND GROUP 2, RESPECTIVELY. MARK X MEANS THAT THE CORRESPONDING FOOD PATCH IS NOT CONSIDERED.

		FP_1	FP_2	FP_3	FP_4	FP_5	FP_6
G_1	Ag 7	X	X	X	0.536	1.000	0.723
	Ag 5	X	X	X	0.524	1.000	0.726
	Ag 2	X	X	X	0.532	1.000	0.725
	Ag 1	X	X	X	0.919	-0.312	1.000
G_2	Ag 6	0.932	0.935	1.000	X	X	X
	Ag 4	0.994	1.000	-0.667	X	X	X
	Ag 3	1.000	0.848	-1.119	X	X	X

The food patch choices for each agent are shown in Fig. 3 for one simulation run. In Group 1, the most dominant agent, Ag 7, is able to forage in the best food patch, FP_5, together with his team members, Ag 5, Ag 2; whereas the weakest agent, Ag 1, must forage in the second best food patch, FP_6. In the second group, Ag 6, 4, and 3 are forced to choose separate patches, although the dominance ranking determines the appropriate food patch as well (with the most dominant agents foraging in the better patches). Fig. 4 provides another simulation result, in which food items in FP_5 are reduced by one and food items in FP_3 increase by one. In both cases, it is clear that the agents seek cooperation for safety in numbers as long as such cooperation is possible (e.g. Ag 7 and 5; and Ag 6 and 4), and compete for food otherwise (e.g. the rest of the agents). The results also reveal that although the current system simplifies the overall foraging problem by decomposing it into manageable smaller problems, suboptimal solutions may occur. For example, in Fig. 3, Ag 3 selected FP_1 rather than FP_4. This suboptimality is mainly due to the grouping algorithm and the initial selection of the area made by the dominant agent. Further refinements of our decision algorithms should help to resolve some of these issues. However, tradeoffs between optimality and computation are likely necessary to solve such complex POSGs and Bayesian games.

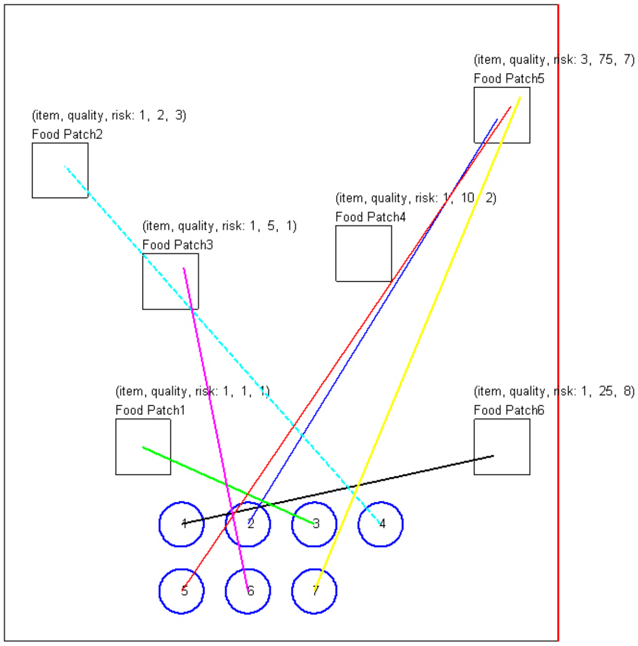

Fig. 3. Case 1: Food patch choices for all agents. Agents 1 to 4 are in the 1st row; agents 5 to 7 are in the 2nd row. The larger the number, the higher the dominance rank.

VI. CONCLUSIONS

This paper examined how agents in a decentralized, multi-agent system make decisions in an environment that is only partially observable and with the decisions of other agents not available. Most existing methods for POSGs limit players to common payoffs. A fundamental disadvantage of this is that they can only be applied to cooperative tasks. We have attempted to utilize game theory to find solutions for general POSGs allowing distinct payoffs. We extended the

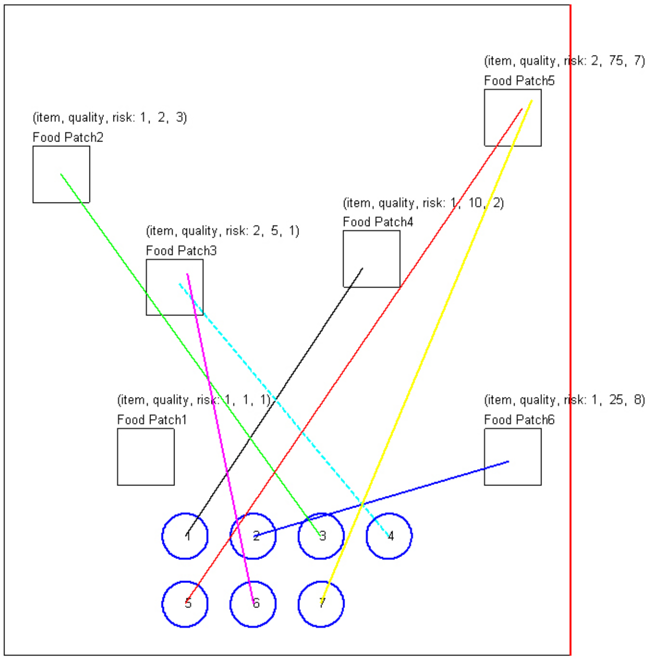
(item, quality, risk: 2, 75, 7)
Food Patch5
(item, quality, risk: 1, 2, 3)
Food Patch2
(item, quality, risk: 1, 10, 2)
Food Patch4
(item, quality, risk: 2, 5, 1)
Food Patch3
(item, quality, risk: 1, 1, 1)
Food Patch1
(item, quality, risk: 1, 25, 8)
Food Patch6
1
2
3
4
5
6
7

Fig. 4. Case 2: Food patch choices for all agents. Compared to Case 1, FP_5 is reduced by one food item while FP_3 is increased by one food item. These food patches can now support one less and one more individual, respectively, which changes the resulting grouping behavior.

Bayesian game approximation to POSGs with distinct payoffs by developing a new heuristic, Q_{MMDP}, the Q values of the corresponding fully observable multi-agent Markov decision process. The simulation results have shown both cooperative and competitive behaviors in partially observable environments.

The computational complexity of such problems might still be considered unsolvable with an increase in the number of players and playing trials. However, social animals have provided an existence proof of their solution to their teamwork tasks. We are therefore concentrating our efforts on examining and modeling the social cognition and behavior of one of the most sophisticated social mammals, chimpanzees. We believe we have taken significant first steps in modeling their behavior in a POSG and Bayesian game framework. Our initial work suggests that several social factors guiding their behavior provide heuristics for significantly reducing the size of the problem, while at the same time encouraging both cooperation and competition. In particular, the biologically inspired grouping method separated the original complex problem into smaller identical problems, and each smaller problem could be analyzed as Bayesian games.

References

[1] D. S. Bernstein, S. Zilberstein, and N. Immerman. The complexity of decentralized control of markov decision processes. *Mathematics of Operations Research*, 27(4):819–840, 2002.

[2] C. Boesch. *Chimpanzee Cultures*, chapter Hunting Strategies of Gombe and Taï Chimpanzees, pages 77–89. Harvard University Press, Cambridge, MA, 1996.

[3] M. Bowling and M. Veloso. Multiagent learning using a variable learning rate. *Artificial Intelligence*, 136(2):215–250, 2002.

[4] R. Emery-Montermerlo, G. Gordon, and S. Thrun. Approximate solutions for partially observable stochastic games with common payoffs. In *Proceedings of the International Joint Conference on Automomous Agents and Multi Agent Systems*, pages 136–143, 2004.

[5] D. Fudenberg and J. Tirole. *Game Theory*. The MIT Press, 1991.

[6] L. P. Kaelbling, M. L. Littmanb, and A. R. Cassandrac. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, pages 99–134, 1998.

[7] A. Kumar and S. Zilberstein. Dynamic programming approximations for partially observable stochastic games. In *Proceedings of the Twenty-Second International FLAIRS Conference*, Sanibel Island, Florida, 2009.

[8] J. C. Mitani, D. P. Watts, and M. N. Muller. Recent developments in the study of wild chimpanzee behavior. *Evolutionary Anthropology*, (11):9–25, 2002.

[9] R. Nair, M. Tambe, M. Yokoo, D. Pynadath, and S. Marsella. Taming decentralized POMDPs: Towards efficient policy computation for multiagent settings. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 705–711, 2003.

[10] F. A. Oliehoek and N. Vlassis. Q-value heuristics for approximate solutions of Dec-POMDPs. In *Proc. of the AAAI spring symposium on Game Theoretic and Decision Theoretic Agents*, page 31C37, 2007.

[11] L. Peshkin, K. E. Kim, N. Meuleau, and L. P. Kaelbling. Learning to cooperate via policy search. In *Proceedings of the Sixteenth International Conference on Uncertainty in Artificial Intelligence*, 2000.

[12] M. Z. Sauter, D. Shi, and J. D. Kralik. Multi-agent reinforcement learning and chimpanzee hunting. In *Proc. of the IEEE International Conference on Robotics and Biomimetics*, 2009.

[13] D. Shi, M. Z. Sauter, and J. D. Kralik. Distributed, heterogeneous, multi-agent social coordination via reinforcement learning. In *Proc. of the IEEE International Conference on Robotics and Biomimetics*, 2009.

[14] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

[15] M. Wooldridge. *An Introduction to MultiAgent Systems*. The Wiley Press, 2009.