A Few Good Agents: Multi-Agent Social Learning

Jean Oh
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
jeanoh@cs.cmu.edu

Stephen F. Smith
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
sfs@cs.cmu.edu

ABSTRACT

In this paper, we investigate multi-agent learning (MAL) in a multi-agent resource selection problem (MARS) in which a large group of agents are competing for common resources. Since agents in such a setting are self-interested, MAL in MARS domains typically focuses on the convergence to a set of non-cooperative equilibria. As seen in the example of prisoner's dilemma, however, selfish equilibria are not necessarily optimal with respect to the natural objective function of a target problem, e.g., resource utilization in the case of MARS. Conversely, a centrally administered optimization of physically distributed agents is infeasible in many real-life applications such as transportation traffic problems. In order to explore the possibility for a middle ground solution, we analyze two types of costs for evaluating MAL algorithms in this context. The quality loss of a selfish algorithm can be quantitatively measured by the price of anarchy, i.e., the ratio of the objective function value of a selfish solution to that of an optimal solution. Analogously, we introduce the price of monarchy of a learning algorithm to quantify the practical cost of coordination in terms of communication cost. We then introduce a multi-agent social learning approach named A Few Good Agents (AFGA) that motivates self-interested agents to cooperate with one another to reduce the price of anarchy, while bounding the price of monarchy at the same time. A preliminary set of experiments on the El Farol bar problem, a simple example of MARS, show promising results.

Categories and Subject Descriptors

I.2.11 [**Distributed Artificial Intelligence**]: Intelligent agents,Multiagent systems,Coherence and coordination

General Terms

Algorithms

Keywords

social learning, coordination, adaptation, price of anarchy, price of monarchy

1. INTRODUCTION

Cite as: A Few Good Agents: Multi-Agent Social Learning, Jean Oh and Stephen F. Smith, *Proc. of 7th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2008)*, Padgham, Parkes, Müller and Parsons (eds.), May, 12-16., 2008, Estoril, Portugal, pp. XXX-XXX.

In this paper, we study distributed learning in a multi-agent resource selection problem (MARS) in which a large number of self-interested agents are competing for common resources. Typically, the goal of multi-agent learning (MAL) in this context is to learn how a particular agent should select resources to maximize its resource utility in the presence of other (possibly also intelligent) agents.

In MAL, two particularly important criteria for learning algorithms are rationality and convergence [5]. The former stresses that a learning algorithm must be adaptive to stationary opponents, while the latter sets the target of convergence to a Nash equilibrium (NE) in self-play - a setting in which opponents are also using the same learning algorithm as the learner. Consequently, the majority of existing MAL algorithms aim for convergence in self-play, particularly to NE [8, 5, 9, 14] or to a set of correlated equilibria (CE) [13, 3].

As seen in the prisoner's dilemma, however, the performance of such non-cooperative equilibria, known as selfish equilibria, can be suboptimal with respect to the natural objective function of a target problem, e.g., in MARS, resource utilization. A mathematical model describing such inefficiency in MARS is referred to as *selfish routing*, of particular interest in both transportation science and computer networks [22].

Although a centrally administered system (CAS) can be used in MARS to guarantee an optimal result, e.g., [6, 4], it is often infeasible in many real-life problem domains such as traffic routing problems in transportation systems. Since agents are physically distributed, the cost of communication among agents and a central administrator is often significant.

In this context, we investigate the possibility of a middle ground solution between selfish equilibria and a centrally administered system wherein self-interested agents utilize a proper level of coordination to improve their performance beyond selfish equilibria.

We first propose the use of two quantitative criteria for evaluating cooperative multi-agent learning algorithms in this context: *price of anarchy*[1] and *price of monarchy*. The former measures the inefficiency of a multi-agent learning algorithm with respect to the natural objective function. Specifically, the price of anarchy is defined as the ratio of the objective function value of a learning algorithm at convergence to that of an optimum. Corresponding to the price

[1]The price of anarchy was originally defined in [16] as the worst ratio of the objective function value of a non-cooperative equilibrium to that of an optimum. We take the liberty to generalize the definition to use it as a criterion for MAL.

of anarchy, we define the price of monarchy, in order to quantify the practical cost of cooperation in terms of communication cost. The quantification of these two criteria provides an eloquent means for trade off analysis for coordination in MARS.

Aspiring to reduce these two costs, we then introduce the A Few Good Agents (AFGA) approach. The key idea of the proposed approach comes from reciprocal altruism. Informally, an agent will cooperate if and only if the agent believes that it would receive higher expected payoffs by taking cooperative actions than by not doing so, and the mutual dependence among agents generates such an incentive for persistent cooperation.

The AFGA approach is driven by a synergy between two types of agents: *leaders* and *voters*. A leader learns to make good predictions about the changes in the environment, and selects actions based on its predictions. When the entire population is composed of agents of a type leader, then the system pays the full price of anarchy.

A voter, alternatively, utilizes *social learning* - learning from other agents – to learn indirectly about the environment from a set of leaders, and selects actions based on its leader's predictions. The intuition for voter agents is similar to that of classical *ensemble learning* methods [10] in that a voter's decisions depend on some set of other learning agents that may use various learning algorithms. The learning algorithms of leaders are transparent to voters, hence, voters' criteria for subscribing to a leader's prediction is determined by the actual performance of the leader's past predictions.

In addition, the AFGA approach presents a beneficial supplementary property. From the perspective of multi-agent systems (MAS), it is redundant to have multiple agents learning the same information simultaneously. We design AFGA agents to utilize social learning to maximize the learning utility of the MAS. This is intuitive if social learning is relatively easier than the learning of the actual task. The name "a few good agents" refers to the subset of agents that are adept learners, namely the leaders, while the other agents - voters – take full advantage of the leaders by only needing to solve a simpler learning problem.

The use of social learning additionally provides robust performance in the case of a dynamically changing population, e.g., cars in traffic routing problems. Most existing MAL algorithms assume a finite set of agents learning at the same speed. Since learning agents tend to explore more at the beginning of the learning phase, the performance of learning algorithms degrades whenever a new set of agents is introduced to the system. The AFGA approach is better insulated from this potential learning degradation caused by dynamic changes in population because new agents can utilize social learning instead of exploring the environment themselves.

We have carried out a preliminary set of experiments in the El Farol bar problem [1], a simple example of MARS, to evaluate the efficacy of our approach. The results show that the AFGA agents significantly reduce the price of anarchy compared to a Nash equilibrium which is a targeted solution of most existing MAL algorithms. At the same time, the AFGA approach has also been shown to bound the price of monarchy significantly lower than that of a centrally administered system.

More notably, the results also demonstrate that the AFGA approach is still effective under two types of uncertainty: 1) when agents have only limited observation of the state of resources, 2) when the members of the agent population in MARS change dynamically, e.g., traffic flow in an automobile routing problem.

2. DEFINITIONS

This section provides a formal definition of MARS, followed by preliminary definitions that are used in our discussion throughout the paper.

2.1 Multi-agent Resource Selection Problem

We formally define MARS as a quadruple of $(N, \Gamma, \overrightarrow{A}, \overrightarrow{R})$. MARS is a single state repeated game which is repeated infinitely.

- N is a set of agents. $N = \{1, 2, ..., n\}$.
- $\Gamma = \{r_1, ..., r_m\}$ denotes a set of resources available for agents in N.
- $\overrightarrow{A}_t = a_1 \times ... \times a_n$ denotes the resource choices of the agents at time t where $a_i \in \Gamma, \forall i \in N$.
- $\overrightarrow{R}_{t+1} : \Gamma \times \overrightarrow{A}_t \rightarrow \Re$ is a delayed reward function

Specifically, a reward associated with using a resource is defined as a function of the number of concurrent users of the resource, and all users using the same resource share the same reward. Thus, MARS is a class of congestion games [20].

2.2 Selfish Equilibria

One of the main criteria of evaluating a MAL algorithm in MARS is convergence. The following two equilibria are often discussed as the target of convergence.

I. Nash equilibrium (NE) [18]: A joint strategy profile π is in a NE iff no player wants to deviate from the current choice of action given the opponents' actions are fixed.

II. Correlated equilibrium (CE) [2]: Given a joint strategy profile π, let π_i, π_{-i} denote the action prescribed for agent i by the strategy π, and the vector of actions prescribed by the strategy π for agents $j \in N, j \neq i$ When a common prior assumption holds, i.e., when all agents have access to the joint distribution of actions of all agents, a correlated equilibrium is a strategy profile s s.t. all players are Bayes rational, i.e., $\forall i \in N, \mathrm{E}[r_i(\pi_i, \pi_{-i})] \geq \mathrm{E}[r_i(a_i, \pi_{-i})] \forall a_i \in A_i$ where A_i denote a set of available actions of agent i.

In fact, a NE is a special class of a correlated equilibrium where agents' decision making (or the information on which an agent's decision depend) is independent from one another.

2.3 Centrally Administered System (CAS)

A centrally administered system (CAS) is a model where a global administrator has access to complete information. An administrator also communicates with all agents in the system, optimizing the performance of the entire population.

3. CRITERIA FOR COOPERATIVE MAL

We propose the use of two criteria to evaluate cooperative MAL algorithms. The quantification of these two values provides important criteria for coordination decisions. Knowingly, agents should coordinate with others if and only if the coordinated actions reduce the price of anarchy while bounding the price of monarchy by a proper level.

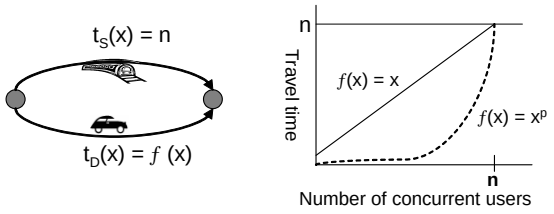

Figure 1: Driving versus Metro

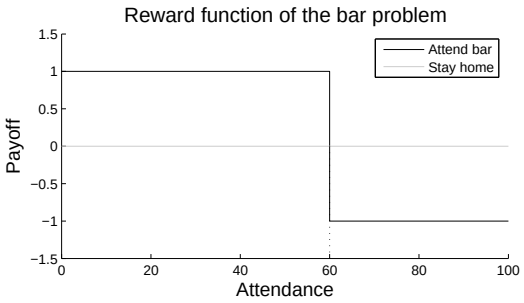

Figure 2: Reward function of EFBP ($n = 100, \tau = 60$)

3.1 Price of Anarchy

The price of anarchy, which was first introduced in [16], measures the inefficiency of a selfish equilibrium. Let φ_a and SE denote the objective function value of a target problem using an algorithm a, and a set of selfish equilibria, respectively. In the price of anarchy literature, it is conventionally assumed that the objective is to *minimize* the cost function φ. Let φ_s and φ_{opt} denote the objective function value of a selfish equilibrium $s, s \in SE$, and that of the optimal solution, respectively. The price of anarchy, $\$^{\mathrm{A}}_{worst}$, is defined as $\max_{s \in SE}(\frac{\varphi_s}{\varphi_{opt}})$. $\$^{\mathrm{A}}_{\mathrm{best}}$ is defined similarly. In the original work, the price is computed using the worst NE. Similar work on correlated equilibria can be found in [7].

In this paper, we generalize the definition such that the price of anarchy measures the inefficiency of a MAL algorithm. Thus, the price of anarchy of a learning algorithm l is

$$\$^{\mathrm{A}}_{\mathrm{l}} = \frac{\varphi_l}{\varphi_{opt}}. \quad (1)$$

3.2 Price of monarchy

Analogous to the price of anarchy, we introduce a new measure, price of monarchy. Whereas the price of anarchy measures potential quality loss due to selfish decisions, the price of monarchy estimates the practical cost of installing cooperation in MAS. We mainly discuss time delay during execution, thus we define the price of monarchy in terms of communication cost.

Thus, the lower bound of the price of monarchy is found in non-communicating systems. In general, the price of monarchy depends on how a coordination mechanism is implemented, e.g., vigorous negotiation may require iterative communication processes. We set the upper bound of the price of monarchy to that of a CAS, and disregard algorithms for which the communication cost exceeds this upper bound.

Let ς_l and ς_{-c} denote a communication cost function of a learning algorithm l, and that of a non-cooperative system, respectively. The price of monarchy, $\$^{\mathrm{M}}_l$, is

$$\$^{\mathrm{M}}_{\mathrm{l}} = \frac{\varsigma_l}{\varsigma_{-c}}. \quad (2)$$

4. SOME BASIC MARS PROBLEMS

In this section, we use two illustrative examples to describe the main issues raised in MARS, namely, 1) the price of anarchy in selfish routing, and 2) the rationality paradox.

4.1 Driving versus Metro

Let us first illustrate selfish routing using a simple example of MARS [19, 22]. Suppose there exist n self-interested agents that are deciding between two actions of taking a metro or driving to work, denoted by M and D, respectively. Figure 1 illustrates such an example. The natural objective of this problem is to minimize the average travel time of all agents.

Let x_a denote the number of agents that selected action a, where $a \in \{M, D\}$, and let $t_a(x_a)$ be the travel time of taking action a. Note that the travel time is a function of x_a, and the agents that selected the same action all experience the same travel time. Let us assume that $t_M(x_M) = n$ where n is the number of agents. On the other hand, let the travel time of driving be a linearly increasing function, $t_D(x_D) = x_D$ such that $t_D(n) = n$. That is, taking a metro takes a constant travel time that is always slower than driving except when the traffic is fully congested.

In this case, self-interested agents converge to an equilibrium in which agents will always choose to drive even when the road is fully congested, resulting the average travel time of n. This is a NE since no one is motivated to deviate from their current choice of actions given the choices of other agents are fixed.

If we assume that there exists a CAS that selects a small number of agents, ϵ, ($\epsilon < n$), and forces them to take the metro, then the travel time of the selected agents is still not any slower than that of their original decisions, i.e., n. This enforcement, however, enables the remaining drivers to travel faster, reducing the travel time of a driving agent to $n - \epsilon$.

In this case, the average travel time of all agents is a convex function, $\frac{1}{n} \times \{n\epsilon + (n-\epsilon)^2\}$. By taking the derivative of the convex function, an optimal value is trivially found, e.g., $\epsilon = \frac{n}{2}$, reducing the average travel time down to $\frac{3n}{4}$.

Suppose instead that the travel time function is non-linear (dotted line in Figure 1), e.g., an exponential function, $t_D(x_D) = x^P$, for some P. Then, in the limit, the travel time of driving becomes ignorable, i.e., $\lim_{P\to\infty}(n-\epsilon)^P = \lim_{P\to\infty}\{n^P(1 - \frac{\epsilon}{n})^P\} = 0$. Thus, the average travel time of agents in this case is reduced to ϵ.

In this example, the price of anarchy is $\frac{4}{3}$ and $\frac{n}{\epsilon}$ for the linear travel time function and the exponential function, respectively. This example provides an interesting observation: although the price of anarchy can be very high, e.g., exponential cost functions, the price can be significantly reduced by a small number of altruistic agents.

4.2 The El Farol Bar Problem (EFBP)

The El Farol bar problem (EFBP) introduced in [1] is

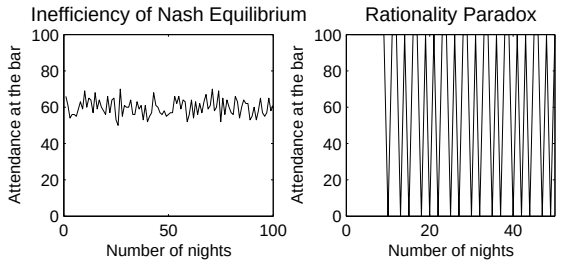

Figure 3: Selfish Equilibria and Rationality Paradox

another example of MARS. EFBP is defined as follows. A set of n agents repeatedly make decisions of whether to attend a bar or not on certain nights. The only observations available to the agents are the past history of attendance at the bar. In the original problem, it was assumed that agents have access to a complete history of attendance.

The payoff of attending a bar is high only if the number of attendees at the bar on the night is less than a certain threshold, τ. However, the agent receives the worst payoff if the bar is over crowded. Thus, an agent is better off staying home if it believes that the bar would be crowded on the night. An example of the reward function for $n = 100, \tau = 60$ is shown in Figure 2.

Despite the simplicity, EFBP clearly exhibits the two important issues of MARS: the *rationality paradox* and *selfish equilibria*. The rationality paradox refers to the fact that a rational agent fails to learn the best action based on its expected reward. Since all agents are simultaneously learning the same information, agents reason in the same manner.

For instance, when an agent predicts the attendance at the bar is lower than τ then the agent decides to attend the bar. Since the other agents also reason the same manner, however, the entire population decides to attend the bar, receiving the worst payoffs. Figure 3 (right) depicts such a result. Agents face contradicting outcomes by making decisions based on their rationality.

Existing studies on EFBP have been mostly focused on the issue of the rationality paradox, seeking algorithms that converge to selfish equilibria. For instance, agents using an inductive reasoning algorithm converge to a mixed strategy NE [1]. Figure 3 (left) depicts a mixed strategy NE of EFBP. Alternatively, agents using a regret-based learning algorithm converge to a set of correlated equilibria [12].

5. A FEW GOOD AGENTS MODEL

In general, some agents can learn better than others at certain times either because they are exposed to different parts of information in the environment, or because they simply have better learning algorithms. Based on this observation, we propose a multi-agent model in which a set of such privileged agents learn directly from the environment while the rest utilize social learning, i.e., they learn rather indirectly from those privileged agents.

We first define two types of agents: *leader* and *voter*. A *leader* agent learns to choose actions for a group of one (itself) or more agents. Instead of learning a policy for itself, an agent may depend on the strategies of other agents. If an agent is following the strategy of another agent it is a *voter*. Thus, a voter agent tries to learn whose strategy yields the highest payoffs.

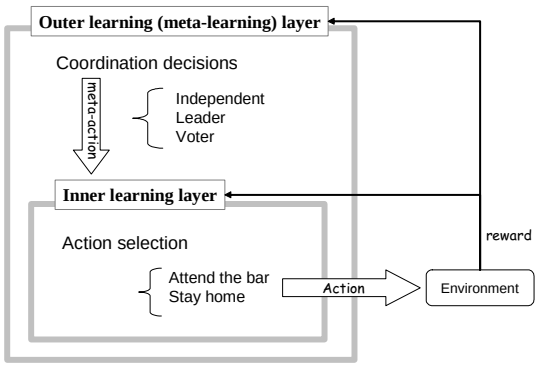

Figure 4: Hierarchical learning layers

An agent can change its type between leader and voter. When an agent is changing its type, it is said that the agent *mutates*. The mutation succeeds with some probability.

Let $type(i)$ denote the type of agent i, $i \in N$. In this context, we hypothesize the existence of a set of leaders at time t, α_t, that can lead the population to perform better than any other set of leaders. Formally,

> Let N be a set of self-interested agents competing for scarce resources, then there exists a set of leaders, $\alpha_t, \alpha_t \subseteq N$ at time t, s.t. $\varphi_t(N, \alpha_t) \leq \varphi_t(N, \alpha'), \forall \alpha' \subseteq N$, where $\varphi_t(N, \alpha')$ denotes the quality loss (in resource utilization) at time t, and $\forall j \in N$, $j \in \alpha'$ iff $type(j) = L$.

That being said, we aim to design learning algorithms for the agent population of MARS to learn to elect the ideal set of leaders themselves.

The crux of the AFGA approach lies in mutual dependence between leaders and voters. On one hand, a leader needs its voters to execute collective actions to achieve a higher payoff, thus it is motivated to be truthful to the voters in order to retain its voting bloc. Since there are multiple leaders, a voter can switch to a new leader if the performance of the current leader is no longer satisfying.

On the other hand, a voter elects to listen to the leader because it believes that acting in the group will produce higher rewards than acting individually. Thus, the AFGA approach is based on reciprocal altruism, which is indeed a class of a selfish model.

The AFGA model is composed of two hierarchical layers as depicted in Figure 4. In general, the goal of reinforcement learning is to learn a policy that maps a state onto an action to maximize the expected future reward over time. The learning of a policy to determine an action, e.g., whether to attend a bar or stay home, is done in the *inner* learning layer. The *outer* layer is where agents learn to choose a coordination action that maximizes the reward. The set of decision choices in the *outer* layer are called *meta-actions*.

As shown in Figure 4, an agent makes decisions in both layers each night, and both the selected meta-action and the ultimate action (attend the bar or stay home) are evaluated according to the same reward received from the environment.

In what follows, we will describe the learning occurring in each layer in detail. Although we will use the bar problem

as an example to illustrate the algorithm, the algorithm is designed for a more general class of a MARS problem.

5.1 Outer learning (meta-learning) layer

In multi-agent systems, an agent needs to decide whether it should act independently or coordinate its actions with other agents in an environment to maximize its performance. The most innovative part of the AFGA approach is representing such coordination decisions as a *meta-learning* layer in the agent's reasoning model.

For instance, an agent learns to make a better decision on whether it should be a *leader* or a *voter* in the outer layer. Alternatively, an agent may still decide to act independently.

The meta-actions representing such choices are denoted by a set: $\{L, I, \alpha_1, ..., \alpha_k\}$. Meta-action L means that the agent is a leader itself whereas meta-action I indicates that the agent is an *independent* voter. An independent voter uses its own policy to select actions. Although a leader without subscribers can be also viewed as an independent agent, it is not included in the set of meta-actions because a leader itself cannot *choose* to become an independent leader in the current model.

Other meta-actions, e.g., α_i, denote candidate leaders. For example, if an agent's meta-action is α_i then the agent is a voter following the strategy of another agent α_i. The maximum number of candidate leaders that one voter can keep in its memory is parameterized by k (default value for k is 4 in the experiments below). Thus, each agent has $k+2$ meta-actions.

Agents learn a policy in the outer layer using a Q-learning algorithm that was simplified for a stateless repeated game [8]. After each night, an agent updates a Q^M-value for its current meta-action. Suppose α_i is the current meta-action. Let r_t be the reward at time t after following the prescription of the current leader α_i. Note that the reward is only defined for the actual action that was taken, e.g., attend the bar, as opposed to meta-actions. The Q^M-value of α_i is then updated as follows.

$$Q^M_{t+1}(\alpha_i) = (1-\eta)Q^M_t(\alpha_i) + \eta r_t \tag{3}$$

where η is the learning rate such that $0 < \eta \leq 1$.

An agent then chooses a meta-action using an ϵ-greedy method, i.e., an agent selects the best meta-action based on the Q^M-values most of time, but it explores other options with a small probability, ϵ.

For instance, a leader agent mutates to a voter if Q^M-value of meta-action L, $Q^M(L)$, is no longer the highest. Conversely, a voter agent mutates to a leader if its $Q^M(L)$ is the highest. With a small probability ϵ, an agent randomly chooses a meta-action rather than the action with the maximum Q^M-value.

5.2 Inner learning layer

In the inner layer, an agent decides the actual action, e.g., whether to attend the bar or not. In fact, every agent has its own stochastic policy π for choosing an action. Initially, the policy of an agent is a random choice among all available actions. The policy is updated when the agent is a leader, and the policy remains stationary when it is a voter.

Depending on its current meta-action choice, an agent may follow its own policy or that of another agent. The strategy of a voter following a leader is straightforward. A voter queries its leader, and the leader uses its policy to prescribe an action for the voter. The voter then simply follows the action that its leader has just prescribed. When a voter is independent, on the other hand, it uses its own (stationary) policy to make a decision.

A more interesting question is how a leader updates its policy. A typical reinforcement learner in the bar problem may try alternative actions and update the value of the two actions - attending the bar or staying home – in order to learn a policy, e.g., *attend* the bar with some probability p.

Instead of updating the value of an individual action of attending the bar, a leader agent updates the expected attendance at the bar. Because the reward in congestion games is defined as a function of the number of the agents that have selected the same action (resource), the expected attendance is in fact the prediction of *joint* actions of the agent population.

Let n denote the number of agents that are deciding to attend the bar, and let τ be the maximum attendance such that agents receive the worst payoffs if the attendance at the bar exceeds τ. The goal of a leader is to regulate the joint actions of its subscribers lest the total attendance exceed the threshold τ assuming the rest of the population is acting selfishly.

Let $l_i(t)$, and $\bar{l_i}(t)$ denote agent i's observation of the actual attendance of the bar on the t^{th} night, and agent i's prediction for the expected attendance on the t^{th} night, respectively. Then the update function is

$$\bar{l_i}(t+1) = (1-\eta)\bar{l_i}(t) + \eta l_i(t) \tag{4}$$

where η is the learning rate such that $0 < \eta \leq 1$.

Based on the expected attendance, a leader further estimates an admissible number of agents among its subscribers that can attend the bar. Since the expected attendance represents the joint actions of the agent population on the next night, the predicted attendance can be interpreted as the sum of two numbers: 1) the number of agents that the leader i is going to send among its subscribers (voters), denoted by $c_{i,t}$, and 2) the number of agents that decide to attend the bar independent of the leader i, denoted by $c_{-i,t}$.

Let $\Omega_{i,t}$ denote a set of voters that subscribe to the leader i's strategy at time t. The length of Ω_i is at least 1 since Ω_i includes leader i itself. An admissible number of agents that can attend on the following night, $c_{i,t+1}$, is estimated by subtracting the load that would be generated by non-subscribers from the predicted attendance $\bar{l}(t)$ as follows.

$$c_{-i,t+1} = \max\{\tau, \bar{l}(t)\} \times (1 - \frac{|\Omega_{i,t}|}{n})\} \tag{5}$$

$$c_{i,t+1} = \max(0, \tau - c_{-i,t+1}) \tag{6}$$

Finally, the policy π is updated as follows. Let $p_{i,t}$ denote the probability that a subscriber of leader i at time t attends the bar.

$$p_{i,t} = \frac{c_{i,t+1}}{|\Omega_{i,t}|} \tag{7}$$

$$\pi_i = (1-\eta)\pi_i + \eta p_{i,t}$$

In other words, policy π_i is the probability that leader i prescribes action *attend* for its subscribers. Since the decision is randomized by policy π, all subscribers receive a fair chance to attend the bar. A leader also counts the number of subscribers for which it has prescribed action *attend* so that the number of attendees among its subscribers at time t is at most $c_{i,t}$ - the admissible number of attendees.

6. EXPERIMENTS

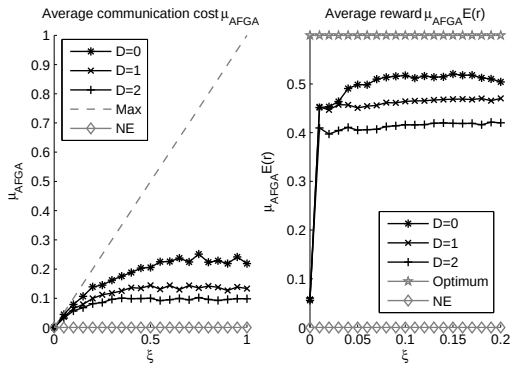

Figure 5: Average reward on varying communication budget

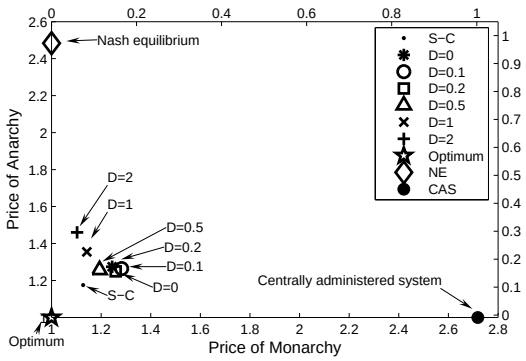

Figure 6: Price curve of EFBP

Prior to reporting the results, we define auxiliary cost functions that were used to compute the price of anarchy and the price of monarchy in the El Farol bar problem (EFBP).

6.1 Auxiliary cost functions

In the El Farol bar problem (EFBP), the objective of a learning agent is to maximize the reward in contrast to the price of anarchy analysis which is commonly conducted in the sense of cost minimization. In order to measure the price of anarchy we define a non-negative, non-increasing function, φ, in terms of the expected reward. The expected reward was normalized to $[0, 1]$.

$$\varphi = 1 - normalize_{[0,1]}(\mu\{E(r)\}) \tag{8}$$

where $\mu\{E(r)\}$ denotes the average of the expected reward of the population. The price of anarchy in EFBP is then computed using Equation 8 and 1.

Subsequently, we define ς of the price of monarchy in EFBP as follows. Let the cost of a message, β, be a non-decreasing function of a communication bandwidth - the number of concurrent users. For instance, the number of concurrent users in a Centrally Administered System (CAS) is n since all n agents need to exchange at least one message with a central administrator at the same time. Thus the average communication cost of a CAS, μ_{CAS}, is

$$\mu_{CAS} = \frac{1}{n} \times \sum_{1}^{n} \beta(n). \tag{9}$$

We now define the communication cost for AFGA. Let V_t denote a set of voters at time t. The cost of a message for a voter at time t depends on the number of other voters that share the same leader. Thus, the average communication cost of AFGA, μ_{AFGA}, is

$$\mu_{AFGA} = \frac{1}{n} \times \sum_{i \in V_t} \beta(|\Omega_{\alpha_{i,t}}|) \tag{10}$$

where $\alpha_{i,t}$ is the current leader of a voter i, and Ω_l denotes a set of voters of a leader l.

Let $\beta(x) = c \times x$ for some constant $c, c > 0$. Without loss of generality, we choose $c = \frac{1}{n}$ such that the communication cost of a non-communicating model, e.g., a Nash equilibrium (NE), and that of a CAS are 0 and 1, respectively. Then we choose ς_l to be

$$\varsigma_l = \exp(\mu_l). \tag{11}$$

We are only interested in learning algorithms for which the price of monarchy $\$^M$ is within the range of $[\$^M_{NE}, \$^M_{CAS}]$.

6.2 Results

We now present a set of experimental results that were performed on EFBP in order to verify the efficacy of the proposed approach. In particular, we discuss the results on the following specific conditions.

I. S-C (Static population, Complete information): This is the original bar problem defined in [1] in which it was assumed that agents have a complete observation on the attendance at the bar.

II. D-I (Dynamic population, Incomplete information): In all conditions except S-C, agents can observe the attendance at the bar only on those nights that they actually attended the bar. In a dynamic population, a subset of the population gradually moves out of the environment, being replaced by new agents. The population dynamics per night is parameterized by D. For instance, $D = 0.1$ denotes that one randomly selected agent is replaced with a new one every 10^{th} night, such that the entire population is substituted every 1000^{th} night when $n = 100$. The set of values used for D in the experiments are $\{0, 0.1, 0.2, 0.5, 1.0, 2.0\}$. Note that $D = 0$ indicates a static population.

All results presented in this section used $n = 100, \tau = 60$, and the results were averaged over 100 trials.

For the purpose of experiment, we introduce an additional parameter for voters: ξ to denote a voter's communication budget, such that communication cost is bounded, $\mu_{AFGA} \leq \xi$. Note that, in general, ξ is set to the upper bound, 1.0, so that each agent makes its own flexible decisions.

We conducted a series of experiments by varying ξ, $\xi \in [0, 1]$. Figure 5 (Left) shows that AFGA agents utilize communication efficiently regardless of the maximum allowance ξ. At the same time, Figure 5 (Right) shows that the performance of AFGA agents starts approaching the optimum even with only a small amount of communication cost.

Since the main goal of this work is to reduce the price of anarchy, the performance of AFGA was compared against that of a mixed strategy NE, which is commonly considered as an *upper* bound performance of MAL, and the optimum (Table 1[2]).

[2]The result on NE is an empirical result.

Table 1: NE vs. Optimum in EFBP

	NE	Optimum
Attendance (μ,σ)	60.0,4.9	60.0,0.0
Average reward	0.007	0.600
$\$_A$	2.483	1.000
$\$_M$	1.000	2.718

In fact, the learning algorithm for a leader in the inner layer converges to a Nash equilibrium (NE) when applied to a strictly non-cooperative setting. By adding the outer learning layer, the AFGA approach allows agents to learn a cooperative policy that improves the performance beyond a Nash equilibrium. In the worst case, if the agents fail to reach a cooperative solution that is better than an independent one, then the agents will learn and act independently, eventually converging to a Nash equilibrium.

In EFBP, the average attendance of NE is 60 ± 4.9, resulting in an average reward near zero. On the other hand, the average attendance of AFGA on all conditions is below the threshold, τ such that $\mu + \sigma \simeq \tau$, reducing the number of overcrowded nights.

Figure 6 presents the price curve of EFBP that summarizes the results by displaying the two cost analysis. We call this a price curve since the price of anarchy measures the quality while the price of monarchy measures the practical cost. The performance of the AFGA approach under 7 conditions were plotted, i.e., each point corresponds to the performance of the AFGA approach at convergence per condition. In addition, Figure 6 also contains 3 baseline performances: the optimum, NE, and CAS. For easier reading, we also plotted scales relative to the price of anarchy of a Nash equilibrium of EFBP, $\$^A_{NE}$, and the price of monarchy of a centrally administered system (CAS), $\$^M_{CAS}$.

The price curve indicates that the price of anarchy is minimized when the system is a CAS by paying the full price of monarchy, and vice versa. Thus, an optimal solution is the lower left corner of the diagram. In general, the closer the data points are to the axes, the better the performance of a learning algorithm is.

The data points in the plot were clustered in the lower left region, which indicates the coordination of AFGA agents is both efficient and effective. In particular, the performance on the original bar problem (S-C) was very close to the optimum. More importantly, the performance was stable under a moderate level of population dynamics, e.g.,$D < 1$ in Figure 6. When the population was highly dynamic, the number of independent agents increased, i.e., the degree of cooperation diminished. Even in such cases, the quality of a solution found by the AFGA approach was significantly higher than that of a NE, e.g., $D = 1$ and $D = 2$ in Figure 6.

7. RELATED WORK

Existing approaches to reducing the price of anarchy in MARS seek solutions from two different sources: one from the environment and the other from the users of the environment. The former addresses making adjustment directly to the environment to make it more efficient. Specific examples are an increase in resource capacity or a redesign of the network routing structure [22]. Our interest resides in the latter, assuming that the environment is not under our control.

The focus of existing MAL has been mostly on convergence to selfish equilibria. For example, there are a number of pure strategy NE that are also optimal in EFBP, e.g., any combination of exactly τ agents attend the bar, while the rest stay home. None of them are, however, stable because they are not fair for those who stay home. Thus, selfish agents converge to a mixed strategy NE which is suboptimal.

Generally in MARS, optimal solutions cannot be achieved without an external intervention or an explicit coordination among agents [17], such that some subset of agents must choose altruistic actions at times. In AFGA, this is accomplished through a leader's strategy for allocating actions among its voters altruistically. In this sense, AFGA can be seen as a local centralization. The approach is, however, unique in that the formation of the organizational structure is dynamically derived by the collective rationality of the agent population as opposed to a predetermined network of organization structure. For instance, AFGA agents mutate themselves between leader and voter according to their performance, which gives rise to self-organizing group behavior. Furthermore, an agent can still decide to be non-cooperative if the expected reward of acting alone is higher.

A no-external-regret algorithm was applied to EFBP in [11]. Their result converged to a set of CE, and the resulting attendance at the bar was 60 ± 5.11. They also proposed a market-based mechanism to achieve fair and optimal solutions for EFBP that bills the bar attendees.

Learning of a periodic policy was introduced in [23] in which agents alternate a set of unfair NE. In their problem domains, it was assumed that agents have access to the performance of other agents. Subsequently, agents act under "homo-egualis" principle, thus social welfare is embedded inside the agents' objective function. For instance, an agent is evaluated not only by its individual performance, but also by the score of the poorest performing agent in the population.

Another common method is to install a centralized control to force a set of agents to take certain actions that are dictated by the administration as opposed to their own choice of actions. Although a completely centralized approach is avoided due to practical reasons, a mixed model of selfish agents and centrally managed agents is commonly used in practice. Virtual Private Network (VPN) is such an example in which intermediate nodes are centrally managed while private users still make independent decisions [15].

The work that is probably the closest to our approach is the Stackelberg strategy [21] in which a set of (market) leaders make moves first, inducing desired responding actions from the followers.

This approach, however, requires leaders to always sacrifice their own payoffs because followers will still choose selfish actions regardless of what moves leaders make. For instance, a Stackelberg strategy performs poorly if a leader adopts a proportional strategy such that it shares the burden only proportionally in the hope that the followers will also share the remaining burden. Thus, a Stackelberg strategy always exploits the centrally controlled set of agents since the followers are not obligated nor motivated to choose altruistic actions.

The hierarchical agent structure of AFGA is similar to the one in Stackelberg games. Regardless of the similarity in its agent structure, our approach is different from the Stackelberg strategy in several major points. First, all agents in the AFGA model make decisions at the same time, in con-

trast to general settings of Stackelberg games. Second, an AFGA leader shares information with its voters a priori as opposed to showing its action choice. Third, AFGA agents self-organize themselves in such a way altruistic actions can be fairly distributed.

8. CONCLUSIONS

This paper makes three contributions. First, we proposed two essential quantitative criteria for MAL: price of anarchy and price of monarchy. Whereas the original definition of the price of anarchy interpreted the quality loss as a coordination cost, we defined a separate measure for the coordination cost to provide an eloquent means for the trade off analysis in MARS.

The second contribution is the introduction of a new learning algorithm, A Few Good Agents (AFGA), that utilizes social learning to reduce the price of anarchy while bounding the price of monarchy. Furthermore, the AFGA algorithm presents a supplementary benefit of maximizing the overall learning utility of MAS through the use of social learning among agents.

Lastly, we demonstrated the usefulness of the AFGA approach in the El Farol bar problem, a special case of MARS with a single resource, and provided a performance analysis in terms of the two quantitative criteria. The results demonstrate that self-interested, rational agents using the AFGA approach learn to reduce the price of anarchy by paying only a small price of monarchy in EFBP. In other words, the AFGA approach avoids both rationality paradox and selfish equilibria. The results also demonstrate that the performance of the proposed algorithm is stable under two types of uncertainty: incomplete observation and dynamic population.

We conclude this paper with a brief note on our future direction on this work. Although the experiment was conducted in a single resource problem, the algorithm was designed for a general class of a MARS problem. Thus, future experiments will be conducted on other general MARS problems.

We also briefly discussed the boosting effect of AFGA with respect to the learning utility of MAS. Our future work includes the definition of a third criterion that quantifies the learning utility of MAS.

9. ACKNOWLEDGEMENT

The authors thank Laura Barbulescu, Anthony Gallagher, and the anonymous reviewers for their valuable comments. This research was sponsored in part by the Department of Defense Advanced Research Projects Agency (DARPA) under contract #NBCHD030010.

10. REFERENCES

[1] W. B. Arthur. Inductive Reasoning and Bounded Rationality (The El Farol Problem). *American Economic Association Annual Meeting, Complexity in Economics Theory*, 1994.

[2] R. J. Aumann. Correlated equilibrium as an expression of bayesian rationality. *Econometrica*, 55(1):1–18, 1987.

[3] A. Blum and Y. Mansour. From external to internal regret. *Journal of Machine Learning Research*, 8(Jun):1307–1324, 2007.

[4] L. Blumrosen and S. Dobzinski. Welfare maximization in congestion games. In *EC '06: Proceedings of the 7th ACM conference on Electronic commerce*, pages 52–61, New York, NY, USA, 2006. ACM.

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

[6] D. Chakrabarty, A. Mehta, and V. Nagarajan. Fairness and optimality in congestion games. In *EC '05: Proceedings of the 6th ACM conference on Electronic commerce*, pages 52–57, New York, NY, USA, 2005. ACM.

[7] G. Christodoulou and E. Koutsoupias. On the price of anarchy and stability of correlated equilibria of linear congestion games. In *ESA2005*, pages 59–70, 2005.

[8] C. Claus and C. Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. In *AAAI/IAAI*, pages 746–752, 1998.

[9] V. Conitzer and T. Sandholm. Awesome: A general multiagent learning algorithm that converges in self-play and learns a best response against stationary opponents. In *ICML*, 2003.

[10] T. G. Dietterich. Ensemble methods in machine learning. *Lecture Notes in Computer Science*, pages 1–15, 2000.

[11] J. Farago, A. Greenwald, and K. Hall. Fair and Efficient Solutions to the Santa Fe Bar Problem. In *Grace Hopper Celebration of Women in Computing*, 2002.

[12] A. R. Greenwald. *Learning to Play Network Games: Does Rationality Yield Nash Equilibrium?* PhD thesis, New York University, 1999.

[13] S. Hart and A. Mas-Colell. A general class of adaptive strategies. *Journal of Economic Theory*, 1:26–54, 2001.

[14] J. Hu and M. P. Wellman. Nash q-learning for general-sum stochastic games. *J. Mach. Learn. Res.*, 4:1039–1069, 2003.

[15] Y. Korilis, A. Lazar, and A. Orda. Achieving network optima using stackelberg routing strategies. *Networking, IEEE/ACM Transactions on*, 5:161–173, 1997.

[16] E. Koutsoupias and C. Papadimitriou. Worst-case equilibria. *Lecture Notes in Computer Science*, 1563:404–413, 1999.

[17] I. Milchtaich. Social optimality and cooperation in nonatomic congestion games. *Journal of Economic Theory*, 114:56–87, January 2004.

[18] J. F. Nash. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, 36(1):48–49, 1950.

[19] A. C. Pigou. *The Economics of Welfare*. Macmillan, 1920.

[20] R. W. Rosenthal. A class of games possessing pure-strategy nash equilibria. *International Journal of Game Theory*, 2(1):65–67, 1973.

[21] T. Roughgarden. Stackelberg scheduling strategies. In *STOC '01: Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 104–113, New York, NY, USA, 2001. ACM Press.

[22] T. Roughgarden. Selfish routing and the price of anarchy (survey). *OPTIMA*, 74, 2007.

[23] K. Verbeeck, J. Parent, and A. Nowé. Homo egualis reinforcement learning agents for load balancing. *Lecture Notes in Computer Science*, 2564:81–91, 2003.