Theoretical Computer Science 411 (2010) 2116–2126

Contents lists available at ScienceDirect

Theoretical Computer Science

journal homepage: www.elsevier.com/locate/tcs

Analyzing the dynamics of stigmergetic interactions through pheromone games

Peter Vrancx [a,*], Katja Verbeeck [b], Ann Nowé [a]

[a] Computational Modeling Lab, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussel, Belgium

[b] Combinatorial Optimization and Decision Support Lab, KaHo Sint-Lieven, Katholieke Universiteit Leuven, Gebroeders Desmetstraat 1, 9000 Gent, Belgium

ARTICLE INFO

Keywords:
Stigmergy
Game theory
Pheromone
Nash equilibrium

ABSTRACT

The concept of stigmergy provides a simple framework for interaction and coordination in multi-agent systems. However, determining the global system behavior that will arise from local stigmergetic interactions is a complex problem. In this paper we propose to use Game Theory to analyze stigmergetic interactions. We show that a system where agents coordinate by sharing local pheromone information can be approximated by a limiting *pheromone* game in which different pheromone vectors represent player strategies. This game view allows us to use established methods and solution concepts from game theory to describe the properties of stigmergy based systems. Our goal is to provide a new framework to aid in the understanding and design of pheromone interactions. We demonstrate how we can use this system to determine the long term system behavior of a simple pheromone model, by analyzing the convergence properties of the pheromone update rule in the approximating game. We also apply this model to cases where multiple colonies of agents concurrently optimize different objectives. In this case a limiting *colony* game can be linked to colony level interactions to characterize the global system behavior.

1. Introduction

The concept of stigmergy was first introduced by entomologist Paul Grassé [1] to describe indirect interactions between termites building a nest. Generally stigmergy is defined as a class of mechanisms that mediate animal to animal interaction through the environment. The idea behind stigmergy is that individuals coordinate their actions by locally modifying the environment rather than by direct interaction. The changed environmental situation caused by one animal, will stimulate others to perform certain actions. This concept has been used to explain the coordinated behavior of termites, ants, bees and other social insects [2].

Recently the notion of stigmergy has gained interest in the domains of multi-agent systems and agent based computing [3–7]. Algorithms such as Ant Colony Optimization (ACO) [8] model aspects of social insect behavior to coordinate agent behavior and cooperation. The concept of stigmergy is promising in this context, as it provides a relatively simple framework for agent communication and coordination. One of the main problems that arises, however, is the difficulty of determining the global system behavior that will arise from local stigmergetic interactions.

In this paper we analyze the dynamics of stigmergetic interactions using game theory. More specifically we examine the case where a colony of agents coordinates its actions by sharing local pheromone information in the environment. We show that the pheromone information in one location can be seen as a strategy of that location in an approximating game played between all locations of the environment. The global system performance is then determined by the payoff achieved in this

[*] Corresponding author.
E-mail addresses: pvrancx@vub.ac.be (P. Vrancx), katja.verbeeck@kahosl.be (K. Verbeeck), ann.nowe@vub.ac.be (A. Nowé).

0304-3975/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.tcs.2010.03.004

Table 1
The Prisoner's Dilemma Game: 2 prisoners committed a crime together. They can either deny their crime (i.e. play the first action) or confess it and betray their partner (i.e. play the second action). When only one prisoner confesses, he gets the maximum reward of 10, while the other one takes all the blame for the crime and receives no reward. When they both deny, each receives a reward of 5, otherwise they only get reward 1.

	a_{21}	a_{22}
a_{11}	(5, 5)	(0, 10)
a_{12}	(10, 0)	(1, 1)

game. The long term behavior of the system can be predicted by analyzing the dynamics of the pheromone update in the limiting *pheromone* game.

In a second stage we extend the model to a situation where multiple pheromone signals and multiple colonies of agents are present in the same environment. In this case agents not only need to optimize their own reward function, but also need to coordinate with other colonies. The resulting problem can be approached at two different levels: one can still look at the pheromone strategies in the approximating game, but it is also possible to look at interactions on the colony level. We will examine the relations between both levels.

Using game theory to analyze learning algorithms is very common in multi-agent reinforcement learning [9–11]. Previously limiting games were used to analyze decentralized learning in Markov games, the general framework for multi-state multi-agent problems [12]. Here the analysis is done from the viewpoint of ant agents and colonies who communicate indirectly via stigmergetic interactions. We directly translate results from multi-agent learning to convergence proofs for ant like algorithms in classical optimization problems.

The remainder of this paper is organized as follows. The next section describes some background material from game theory that will be used later on. Section 3 describes the model of stigmergetic interactions similar to that of ACO algorithms. Results for the single-colony setup and multi-colony setup are derived. These theoretical results are experimentally demonstrated in Section 5.1 on a simple biologically inspired optimization problem and on a small routing instance, a successful ACO application. We conclude this paper with a final discussion.

2. Game theory

The central concept of game theory is to model strategic interactions as a game between a set of players. In this section we review basic game theoretic terminology and define two common solution concepts in games: Nash equilibria and Pareto optimality. A detailed overview of normal form games and their solutions can be found in [13,14].

Assume a collection of n agents where each agent i has an individual finite set of actions A^i. The number of actions in A^i is denoted by $|A^i|$. The agents repeatedly play a single stage game in which each agent i independently selects an individual action a from its private action set A^i. The combination of actions of all agents at any time step, constitute a joint action or action profile $\vec{a}$ from the joint action set $\mathbb{A} = A^1 \times \cdots \times A^n$.

With each joint action $\vec{a} \in \mathbb{A}$ and agent i a distribution over possible rewards is associated, i.e. $R^i : \mathbb{A} \to \mathbb{R}$ denotes agent i's expected payoff or expected reward function.

The tuple $(n, \mathbb{A}, R^{1\ldots n})$ defines a single stage strategic game, also called a normal form game. In the two player case, normal form games are often represented by their payoff matrix. An example of this can be seen in Table 1. In this case the action selected by player 1 selects a row in the matrix, while that of player 2 determines the column. The corresponding entry in the matrix then gives the payoff received for that play. Players 1 and 2 are also referred to as the row and the column player, respectively.

A *strategy* $\sigma^i : A^i \to [0, 1]$ is an element of $\mu(A^i)$, the set of probability distributions over the action set A^i of player i. This strategy is called a pure strategy if $\sigma^i(a^i_j) = 1$ for some action j and 0 for all other actions, otherwise it is called a *mixed strategy*. A strategy profile $\vec{\sigma} = \{\sigma^1, \ldots, \sigma^n\}$ is a vector of strategies, containing one strategy for each agent.

At each time step $t = 0, 1, 2, \ldots$ all agents individually select an action from their action set, resulting in the play $a(t) = \{a^1(t), \ldots, a^n(t)\}$. After each interaction the agents are informed of the payoff they have obtained during this round and have the possibility to update their strategy before the next round starts.

Definition 1. Let $\vec{\sigma}^{-i}$ denote the profile $\vec{\sigma}$ minus the strategy of agent i. A strategy profile $\vec{\sigma}$ is then called a Nash equilibrium when for all i we have:

$$R^i(\vec{\sigma}) \geq R^i(\vec{\sigma}^{-i} \cup \{\sigma\}) \quad \forall\, \sigma \in \mu(A^i).$$

A Nash equilibrium $\vec{\sigma}$ is said to be pure when all strategies in $\vec{\sigma}$ are pure. Nash proved [15] that every strategic game, with finitely many actions has at least one (possibly mixed) Nash equilibrium. From the definition it is clear that in a Nash equilibrium each agent is playing a best response to the current strategies of the other players. This means that, no player has an incentive to unilaterally deviate from this strategy profile.

Another well-known solution concept is given by the notion of *Pareto optimality*:

Definition 2. A strategy profile $\vec{\sigma}_1$ is said to be Pareto optimal if there is no other strategy profile $\vec{\sigma}_2$ in which all players simultaneously do better and at least one player is doing strictly better. The set of Pareto optimal strategies is called the Pareto front.

Note that a Nash equilibrium is not necessarily Pareto Optimal and vice versa, a Pareto optimal solution is not necessarily Nash. This can be seen in the Prisoner's Dilemma Game of Table 1. The Nash equilibrium in this game is for both players to confess (i.e. betray each other) and receive a reward of 1. This equilibrium is not Pareto optimal, however, as both players could do better if they both denied the crime.

3. A single-colony model of stigmergetic interactions

In this section we describe our stigmergetic model, which was originally introduced by Verbeeck and Nowé [6]. The purpose of this model is to capture the essence of the pheromone dynamics in systems similar to Ant Colony Optimization algorithms. To this end we include only the basic elements of stigmergetic coordination. The model abstracts pheromone communication as a set of location based strategies which are shared among multiple agents. Currently, the model is limited to pure pheromone interactions and additional optimization tools such as heuristic information and local search improvements are not modeled. The description given here is based on the Markov Decision Process (MDP) formalization [16]. This setting is popular in stochastic optimization but can also be used to describe certain discrete optimization problems [17].

We consider an environment consisting of a set of discrete locations $L = \{l_1, \ldots, l_n\}$. Each location l has a set of possible outgoing links that can be followed, which we alternatively call actions $A^l = \{a^l_1, \ldots, a^l_r\}$ that can be performed in that location. Further we have a transition function $T(l, l', a)$, which gives the probability to move from location l to l' when taking action $a \in A^l$ and a reward function $R(l, l', a)$ which gives the reward for this transition.

In this environment a set or *colony* of ant agents learns a policy π which maps each location $l \in L$ to an action $a^l \in A^l$. The goal of the colony is to learn a policy which maximizes the average reward over time:

$$J^\pi \equiv \lim_{N\to\infty} \frac{1}{N} E\left[\sum_{t=0}^{N-1} R^\pi(l(t), l(t+1))\right]. \tag{1}$$

Ants belonging to the same colony collaboratively learn a single policy. To do this, they base their action selection in location l on a shared pheromone signal τ^l that associates a value τ^l_i with each action $a^l_i \in A^l$. A pheromone learning system then consists of 2 components: an update rule which governs changes in pheromones based on received reinforcements and a normalization which is used to generate action probabilities from the pheromones. For a location l these probabilities are determined from the local pheromone signal τ^l, by applying the normalization function g, so that $Pr\{a(t) = a\} = g(a, \tau)$.

As is common in pheromone based systems, we update local pheromones using a reward signal over an entire episode, rather than using the immediate reward $R(l, l', a)$. More precisely, when an ant returns to a previously visited location, it updates the local pheromones using the following estimate β of the global average reward:

$$\beta = \frac{\Delta r}{\Delta t}. \tag{2}$$

Here Δr is the reward gathered since the last visit, and Δt is the number of time steps since this last visit.

In this paper we focus on a system were τ^l is a probability vector and $g(a^l_i, \tau^l) = \tau^l_i$. This means that ant agents directly update the action probabilities and we have for all locations l: $\sum_j \tau^l_j = 1$ and $0 \leq \tau^l_j \leq 1$ $\forall j$. To update the values we use the so called linear reward-inaction (L_{R-I}) scheme from learning automata theory [18]:

$$\tau_i \leftarrow \tau_i + \lambda\beta(1 - \tau_i) \tag{3}$$
$$\text{if } a_i \text{ is the action taken}$$
$$\tau_j \leftarrow \tau_j - \lambda\beta\tau_j \tag{4}$$
$$\text{if } a_j \neq a_i$$

where we assume that the reinforcement β has been normalized to lie in $[0, 1]$. The constant $\lambda \in [0, 1]$ is called the *learning rate* and determines the influence of pheromone deposit β. It is similar to the evaporation rate often used in pheromone based systems.

The update system describe above conforms to the traditional idea behind pheromone updates that the probability of an action is increased relative to the quality of solutions in which it was used. The similarities between L_{R-I} and the update used in Ant System [19] are discussed in [6,20]. One advantage of using this scheme is that convergence results for a wide range of possible settings are available. For instance in [21,22] it is shown that L_{R-I} can converge to a pure Nash equilibrium in repeated games, provided that the learning rate λ is small enough.

The model described here was used in [6] to show optimality for the pheromone system with L_{R-I} update, based on existing results from MDP literature [23]. In the next subsection we will show how this model can be used more generally, to analyze a pheromone update scheme in terms of their behavior on a strategic game which approximates the optimization problem under study. We will then continue by extending the model to allow for multiple colonies of agents, and demonstrating the game theoretic analysis that is possible in this extended case.

3.1. Analysis of the single-colony model

We now show how the use of the model above allows us to approximate the behavior of the pheromone system by a repeated strategic game. This approach was used previously in [23] to show the convergence of an algorithm for MDPs using L_{R-I} and in [12] in the context of multi-agent reinforcement learning. The stigmergy model described in the previous section allows us to apply the same analysis to pheromone systems.

A critical assumption we need to make here is that the Markov chain of locations generated by a single ant agent is ergodic under all possible policies. This means that under any fixed policy π all locations continue to be visited and updated. Additionally, the process converges to a stationary distribution over the locations and for each policy π we have a probability distribution d^π over the locations:

$$d^\pi = \{d^\pi(l_1), \dots, d^\pi(l_n)\} \quad \text{with} \sum_{l \in L} d^\pi(l) = 1, \quad \text{and} \quad d^\pi(l) > 0, \forall l$$

where $d^\pi(l)$ represents the probability of the ant agent being in location l. This probability is independent of the time step and starting location. The assumption of ergodicity is necessary to ensure that pheromone strategies in all locations keep being updated. It is a common assumption in average reward reinforcement learning [24,23]. It also enables the approximation of asynchronous pheromones updates by a synchronous repeated strategic game. More information on this approximation can be found in [23].

Under this ergodicity assumption, it is possible to approximate the single-colony model by a game as follows. Consider all locations $l \in L$ of the environment as a players. The action set for each player is exactly the action set A^l of the corresponding location. The pheromone vector τ^l (together with function g) represents the current strategy for this player. Since we have one player for each location, a play in this game maps every location to an action and represents a pure policy π over all locations. The payoff that players receive for this play in the game is the expected average reward J^π for the corresponding policy π. Using the stationary distribution over the locations d^π this reward can be written as follows:

$$J^\pi = \sum_{l} d^\pi(l) \sum_{l' \in L} T^\pi(l, l') R^\pi(l, l') \tag{5}$$

where T^π and R^π are the expected transition probabilities and rewards under policy π. As each play (or policy) gives the same reward for all players the game is called an identical payoff game or a team game.

The game approximation described above was introduced in [23] to prove the convergence of L_{R-I} in average reward MDPs. Following result was proved for the approximating game obtained using the method above:

Theorem 1 (*Wheeler, 1986*). *Let $\Gamma = (n, A, J)$ denote the n-player identical payoff game where player l has action set A^l and a play $\vec{a} \in A^1 \times \cdots \times A^n$ receives payoff $J^{\vec{a}}$ as defined in Eq. (5). Then this game has a unique pure equilibrium, corresponding to the policy that maximizes J.*

Since the L_{R-I} update scheme was proven to converge to pure Nash equilibria in repeated games [22,23], the above theorem guarantees that the players will converge to the optimal policy, i.e. the policy that maximizes J. In [6] Verbeeck and Nowé showed that these results still hold in the case of multiple cooperative ant agents updating the same action probability vectors. In this setting ant agents are responsible only for sampling the reward and triggering updates. The actual learning and intelligence is stored in the pheromone vector update in each location $l \in L$. Thanks to the ergodic assumption all locations will continue to be updated, and these asynchronous ant updates can be approximated by synchronous learning in the constructed game.

Of course the above model can be used with other pheromone update systems. The limiting games described above, depend only on the problem and not on the pheromone update used. Any combination of a local pheromone update with a normalization to action probabilities could be treated as a learning strategy in this approximating limiting game. In order to predict the outcome such a system will obtain, we need to study its dynamics on the approximating game. As an example, we examine the pheromone update in Eq. (6), which is used in Ant Colony System [8]:

$$\tau_i \leftarrow (1-\lambda)\tau_i + \lambda\beta \qquad \text{if } a_i \text{ is the action taken.} \tag{6}$$

We shall examine the dynamics of this system when used with a Boltzmann normalization:

$$g(a_i^l, \tau^l) = \frac{e^{\tau_i^l/T}}{\sum_{a_j^l \in A^l} e^{\tau_j^l/T}}. \tag{7}$$

This distribution function assigns each action a probability based on the associated pheromone value and a parameter T, called the temperature. This parameter determines the amount of influence a difference in pheromones has on the action probabilities. Higher values cause actions to be close to equiprobable and lower values result in greater differences in probabilities. The repeated game dynamics of the update in Eq. (6) when used with a Boltzmann normalization are studied in [10]. In Section 5.1 we give an example of the behavior of this system, as an alternative to the L_{R-I} update.

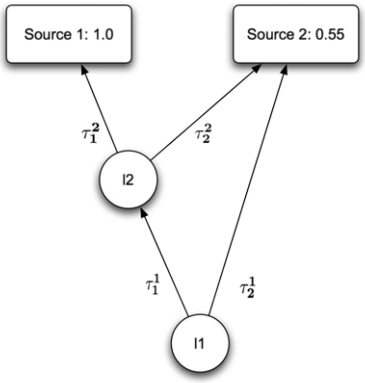

Fig. 1. Example of a single-colony problem. A foraging ant colony must decide which food source to exploit.

Table 2
Approximating game for the problem shown in Fig. 1. Players correspond to problem locations. The game payoffs are the expected averages over time of the amount of food collected under the corresponding policies. The unique equilibrium is shown in bold.

		l2	
		a1	a2
l1	a1	**0.32**	0.20
	a2	0.28	0.26

As an example consider the very simple, biologically inspired optimization problem depicted in Fig. 1. A colony of ant agents is trying to optimize the rate at which they collect food. Starting from their nest at location l_1 they can collect food from 2 sources with source 1 having an expected payoff of 1.0 units and source 2 giving an average of 0.55 units of food. From location l_1 they can choose either to proceed to location l_2 (action 1) or to go directly to source 2 (action 2). From location l_2 they can go to either source 1 (action 1) or source 2 (action 2) directly. Transitions are assumed to be stochastic with ants having a 0.9 probability of arriving at the chosen location and a 0.1 probability to make the transition associated with the other action. When an ant agent reaches a food source it receives a payoff depending on the food source and returns to the nest at location l_1. The agents have to decide whether to exploit the closer food source or to concentrate on the richer but also further source.

The game approximating the problem above is shown in Table 2. Since we have 2 locations with 2 possible actions each, the game is a 2 player, 2 action game. The row and column actions here are the possible actions in location l_1, l_2, respectively. The game has 1 pure Nash equilibrium corresponding to the policy which plays action 1 in both locations and thus prefers food source 1. This equilibrium indeed corresponds to the optimal policy. Experimental results on this problem are provided in Section 5.1.

4. A multi-colony model of stigmergetic interactions

In this section we extend the previous model by adding ant agents belonging to different colonies. In a multi-colony system, each colony uses its own pheromone signal, where pheromones are now only shared by ants belonging to the same colony. Different colonies correspond to different objectives, represented by different reward functions, and are therefore suited to solve multi-objective optimization problems. It should be noted that our model is mainly based on the routing and load-balancing applications described below, rather than the combinatorial multi-objective problems treated for example in [25]. The model described here was inspired by the AntNet routing algorithm [26]. Here ant agents in the system have different goals, as they need to find routes to multiple destinations. Therefore, AntNet uses different pheromones, corresponding to the possible destination nodes. Ants finding routes to the same destination can be seen as a separate colony, sharing pheromone information with each other, but not with ants that have a different destination. However, the colonies in AntNet cannot simply be treated as several single colonies learning in parallel. Since all colonies use the same network links, the performance of a colony is influenced by the strategies of other colonies. Network routes that are heavily used by other colonies will become less attractive, since they can suffer higher delays. As such, to analyze these types of multi-colony systems we also need to model interactions between colonies. We now give a description of this extended model, which is based on the framework of Markov Games [9], the classical extension of MDPs to multi-actor environments.

Consider m objectives and thus m colonies present in the problem. The environment still consists of a set of discrete locations $L = \{l_1, \dots, l_n\}$ with a set of outgoing links or actions $A^l = \{a_1^l, \dots, a_r^l\}$ that can be followed.[1] However, in each

[1] For ease of notation we assume that each colony has the same action set in every location.

location l, m different pheromone vectors $\tau^{l,c}, \;\; c : 1 \ldots m$ are now present, representing the action probabilities for an ant agent of colony c in location l. Transitions to new locations can still be defined by $T(l, l', a)$ as being the probability of one ant going to location l' by taking action a in location l independent of colony type c. As before, the pheromone probability vectors $\tau^{l,c}$ are updated by ant agents walking around in the environment using the reward gathered over the entire episode as specified by Eq. (2). However, each colony c now has a different reward function, the value of which can be influenced by the behavior of another colony. The expected average reward for a single colony now depends on the joint policies $\Pi = \{\pi_1, \ldots, \pi_m\}$ of all colonies. Denote Π_{-c} for the joint policy of all colonies minus the policy of colony c, then $R^c(l, l', a, \Pi_{-c}), c : 1 \ldots m$ specifies the reward for an ant agent of colony c, taking action a in location l, moving to location l' with Π_{-c} representing the interference of ant agents not in colony c. The goal of each colony c is to learn a policy π_c which maximizes the average reward over time:

$$J^{\Pi,c} \equiv \lim_{N\to\infty} \frac{1}{N} E\left[\sum_{t=0}^{N-1} R^{\pi,c}(l(t), l(t+1), \Pi_{-c})\right]. \tag{8}$$

Since the problem is multi-objective, optimizing for one colony, might lead to lower rewards for other colonies. In the next subsection we explain that this setting can now be analyzed in terms of the colonies' behavior on two different strategic games.

4.1. Analysis of the multi-colony model

Again we make an ergodicity assumption. Consider one ant agent from colony c moving in the environment with other ant agents from all other colonies excluding colony c. We assume that the Markov chain generated by this ant agent is ergodic, meaning that again a stationary distribution over the locations exist and the average reward over time for colony c can be calculated as follows:

$$J^{\Pi,c} = \sum_{l} d^{\pi,c}(l) \sum_{l' \in L} T^{\pi}(l, l') R^{\pi,c}(l, l', \Pi_{-c}) \tag{9}$$

with $d^{\pi,c}(l)$ the stationary distribution of one ant agent of colony c over the locations under policy π. The resulting problem can now be viewed at two levels. Instead of making the locations the players of the game, we can look at the game between the different pheromone strategies, i.e. the pheromone vectors. Since m vectors are present in each location, there are m players for each location. A play in this game results in an action for each colony in every location and thus represents a pure joint policy. The game will no longer be a identical payoff game since players belonging to different colonies follow different reward functions. Players within one colony c do receive the same payoff $J^{\Pi,c}$ (given by Eq. (9)) for a play Π. In total we will have a game consisting of $n \times m$ players, with n the number of locations and m the different number of colonies, each having r actions. We refer to this game as the *pheromone* game, since each player represents a pheromone strategy in a single location.

At another level we look at the expected rewards resulting from the interaction between the colonies. We do this by considering all the possible pure policies for the colonies as actions in a single large game. In this game the players are the colonies and the payoff is the expected reward a colony obtains for a certain combination of policies from all colonies. In total we have now a game of m players each having $r \times n$ actions. We refer to this game as the *colony* game.

From the description above it is clear that each play in both the pheromone and the colony game maps an action to each location for every colony. Thus a play in either game represents a pure joint policy over all locations. Furthermore for every play in the colony game we have a corresponding play in the pheromone game and vice versa. We can now show following relation between both games:

Theorem 2. *Let Γ_1 denote the colony game were each player (or colony) c has action set $A^1 \times \cdots \times A^n$ and receives payoff $J^{\vec{a},c}$ for play $\vec{a} \in (A^1 \times \cdots \times A^n)^m$. Let Γ_2 denote the pheromone games, where every player (or pheromone vector) in location i has action set A^i and receives payoff $J^{\vec{a},c}$ for a play $\vec{a} \in \prod_{i:1\ldots n}(A^i)^m$. Then a joint action $\vec{a}$ is a pure Nash equilibrium of Γ_1 if and only if the corresponding joint action $\vec{a}$ is a pure Nash equilibrium of Γ_2.*

Proof outline:

if: In a Nash equilibrium no player can improve his payoff by unilateral deviation of the current strategy profile. So in a Nash equilibrium of the colony game, no player has a policy which gives a higher payoff provided that the other players stick to their current policy. A player (or pheromone vector) of colony c switching its action in the pheromone game corresponds to a single player (here colony c) in the colony game switching to a policy which differs in just 1 location. As the players representing pheromone vectors belonging to colony c in the pheromone game receive the same payoff as the corresponding colony c in the colony game, the pheromone player cannot improve its payoff.

only if: Suppose a play Π exists which is a Nash equilibrium in the pheromone game, but the corresponding play is not a Nash equilibrium in the colony game. This means that in the colony game we can find at least one player or colony c which can switch from its current policy π_c to a better policy π_c' while other players keep their policy constant. But the situation where other colonies keep their policies fixed corresponds to a 1 colony problem without any external influences for the reward function. So for colony c where the other colonies play their fixed strategy Π_{-c}, the reward

Table 3
Colony game approximating the multi-colony version of Fig. 1. Equilibria are indicated in bold.

		Colony 2			
		(a1, a1)	(a1, a2)	(a2, a1)	(a2, a2)
Colony 1	(a1, a1)	0.235, 0.235	0.295, 0.184	**0.288, 0.254**	0.297, 0.246
	(a1, a2)	0.184, 0.295	0.149, 0.149	0.128, 0.205	0.122, 0.183
	(a2, a1)	**0.254, 0.288**	0.205, 0.128	0.176, 0.176	0.169, 0.152
	(a2, a2)	0.246, 0.297	0.183, 0.122	0.152, 0.169	0.142, 0.142

function is given by taking the expectation of reward R with respect to the fixed Π_{-c} and the resulting expected average payoff is given by $J^{\pi_c} = J^{(\Pi_{-c},\pi_c),c}$. According to Theorem 1 this situation can be represented by a game Γ with a unique, optimal equilibrium. Since $J^{(\Pi_{-c},\pi_c),c} < J^{(\Pi_{-c},\pi_c'),c}$, π_c cannot be this equilibrium and a policy must exist which achieves a higher payoff but differs in only 1 location.[2] But since this policy would also receive a higher payoff then π_c in the pheromone game, $\Pi = (\Pi_{-c}, \pi_c)$ cannot be a Nash equilibrium of this game which leads to a contradiction.

The relation between both games allows us to predict colony behavior based on the convergence properties of the local pheromone update in the pheromone game. For instance, from the theorem above and the pure equilibrium convergence of L_{R-I} we immediately get following result:

Corollary 1. *Consider an optimization problem with location set L and C colonies optimizing individual reward functions. If all colonies use the L_{R-I} update with a sufficiently small learning rate, the system will converge to a pure Nash equilibrium between the policies of the colonies.*

We demonstrate these results on an extended version of the example in Fig. 1. Instead of a single colony optimizing its food collection, we now consider 2 colonies gathering food in the same environment. Each colony needs to optimize its own foraging, but also has to coordinate with the other colony. When both colonies select the same food source, they have to share and thus will receive a lower payoff. So instead of a fixed average payoff for each source, an ant reaching a food source now receives a payoff which is also based on the number of ants from other colonies at that source. More specifically when arriving at a food source we give the ant a reward of $(1.0 - \frac{q_s}{q_{tot}}) * p_s$ where q_s is the number of ants from other colonies at the food source, q_{tot} is the total number of ants from other colonies in the system and p_s is the average payoff for the source (i.e. 1.0 and 0.55 for source 1 and 2, respectively).

In Table 3 we show the colony game for the 2 colony problem described above. Since we have 2 colonies this game has 2 players. The actions for these players are the possible pure policies for the corresponding colony. Since both colonies have 2 locations with 2 possible actions, they have 4 pure policies or actions in the game. The payoffs for the game are the expected average payoffs for the resulting joint policy, as defined in Eq. (9). Since these payoffs differ for the colonies the approximating game is not a team game, as was the case in the single-colony problem. Payoffs in the table are calculated by calculating the stationary distribution of locations corresponding to the different policies (see for example [16] for a detailed explanation) and using these values in Eq. (9).

The 2 pure Nash equilibria are indicated in bold. These equilibria correspond to a situation where 1 colony exploits source 1, but the other one selects source 2 in location 1. It should be noted that these equilibria give different payoffs for the colonies, with one colony receiving an average payoff of 0.288 and the other one receiving only 0.254. This results in different preferences of the colonies for both equilibria. Furthermore, the equilibria do not give the maximum possible reward for either colony, but they are Pareto optimal.

The pheromone game corresponding to this problem is shown in Table 4. This game consists of 4 players $p^{l,c}$, corresponding to the pheromone vectors $\tau^{l,c}$. Since two actions are present in each location, all players have two possible actions. Players corresponding to the same colony receive the same payoff for a play. These payoffs also correspond to the values achieved by the colonies in the game of Table 3. Equilibria of the game are indicated in bold. It can easily be verified that these equilibria correspond to the same joint policies as those of the colony game.

5. Experiments

5.1. A simple example

We first demonstrate the single-colony case. Fig. 2(a) shows the average reward over time obtained by a single colony using the L_{R-I} update. To demonstrate the convergence we do not show an average over multiple runs, but rather a single typical run where we allow the average reward to converge and then randomly reinitialize the pheromone vectors. The figure shows the average reward over time obtained by a single ant of the colony at each time step, for a total of 25 000 time

[2] If such a policy did not exist π_c would be an equilibrium of Γ.

Table 4
Pheromone game approximating the multi-colony version of Fig. 1. Column 1 lists possible plays, with column 2 giving the expected payoff for each player resulting from a play. Equilibrium plays are indicated in bold.

$(p^{1,1}, p^{1,2}, p^{2,1}, p^{2,2})$	$J^{\Pi,1}, J^{\Pi,2}, J^{\Pi,1}, J^{\Pi,2}$
(a1, a1, a1, a1)	0.235, 0.235, 0.235, 0.235
(a1, a2, a1, a1)	**0.288, 0.254, 0.288, 0.254**
(a1, a1, a2, a1)	0.184, 0.295, 0.184, 0.295
(a1, a2, a2, a1)	0.128, 0.205, 0.128, 0.205
(a2, a1, a1, a1)	**0.254, 0.288, 0.254, 0.288**
(a2, a2, a1, a1)	0.176, 0.176, 0.176, 0.176
(a2, a1, a2, a1)	0.246, 0.297, 0.246, 0.297
(a2, a2, a2, a1)	0.152, 0.169, 0.152, 0.169
(a1, a1, a1, a2)	0.295, 0.184, 0.295, 0.184
(a1, a2, a1, a2)	0.297, 0.246, 0.297, 0.246
(a1, a1, a2, a2)	0.149, 0.149, 0.149, 0.149
(a1, a2, a2, a2)	0.122, 0.183, 0.122, 0.183
(a2, a1, a1, a2)	0.205, 0.128, 0.205, 0.128
(a2, a2, a1, a2)	0.169, 0.152, 0.169, 0.152
(a2, a1, a2, a2)	0.183, 0.122, 0.183, 0.122
(a2, a2, a2, a2)	0.142, 0.142, 0.142, 0.142

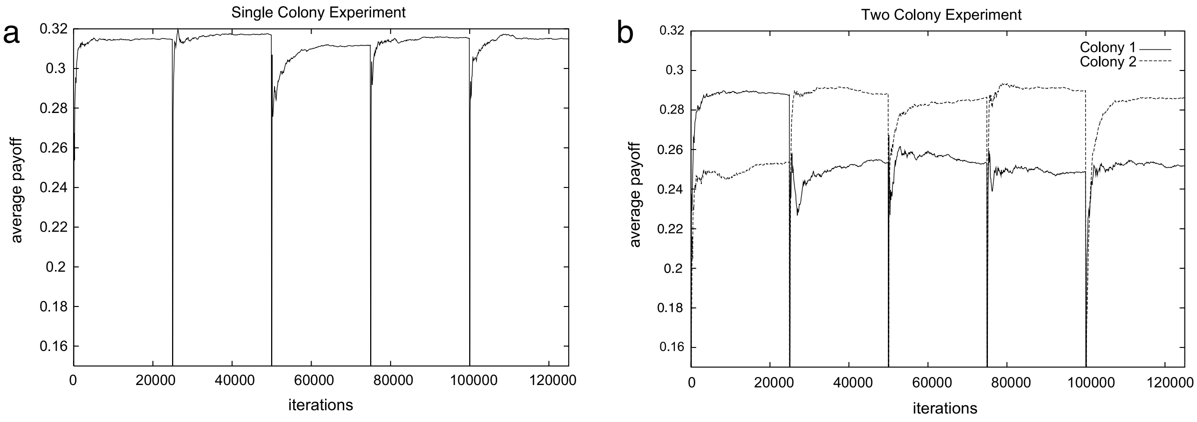

Fig. 2. Results for L_{R-I} update on the example of Fig. 1. (a) Single-colony experiment. (b) Two colony experiment. Settings where $\lambda = 0.001$ and 100 ants per colony.

steps for each initialization. A single time step here corresponds to the selection of 1 action by all ants in the system, i.e each ant makes 1 transition between locations. In each case the system converged to the predicted equilibrium (1, 1) and after 25 000 time steps the average reward approached the predicted value to within 0.005.

Results for the two colony problem described in Section 4 can be seen in Fig. 2(b). Here the system converged to either one of the equilibria in Table 4, with one colony receiving an average payoff of 0.254 and the other one receiving 0.288. Again the obtained values after 25 000 iterations can be seen to closely approximate the predicted values. The eventual equilibrium reached depends on the initialization of the pheromone vectors.

Additionally, we give an example of how this analysis can be used with another pheromone update system. We again apply this update on the problem in Fig. 1. Since this is the same problem as studied above, the approximating limiting game is still the game given in Table 2. However, we will demonstrate that a different pheromone update can give very different outcomes. We give results for the pheromone update in Eq. (6), together with a Boltzmann normalization. In [10], the learning dynamics for this scheme are studied, by determining a continuous time approximation using ordinary differential equations (ODEs). It is shown that this continuous time limit of the dynamics is an adapted version of the continuous time replicator dynamics from evolutionary game theory [13]. Using these ODEs, we can predict the outcome of the update on a strategic game by determining the stationary points of the system and investigating their stability. A similar approach was used in [22] to show the equilibrium convergence of the L_{R-I} update in strategic games.

Rather than explicitly calculating the stationary points for the system, we visualize the dynamics using a direction plot. Fig. 3(a) shows a plot of the predicted evolution of the action probabilities for both locations when using a Boltzmann normalization with $T = 0.2$. This plot was obtained by analyzing the dynamics of the replicator equations on the approximating game in Table 2. The visualization immediately shows that for these settings the update scheme will not reach the optimal pure equilibrium, but rather will go to a mixed outcome. Fig. 3(b) shows the results for multiple runs with different pheromone initializations on the single-colony version of the problem in Fig. 1. The ants used a Boltzmann function with $T = 5$ for action selection and the pheromone update in Eq. (6) with $\lambda = 0.001$, with the colony consisting

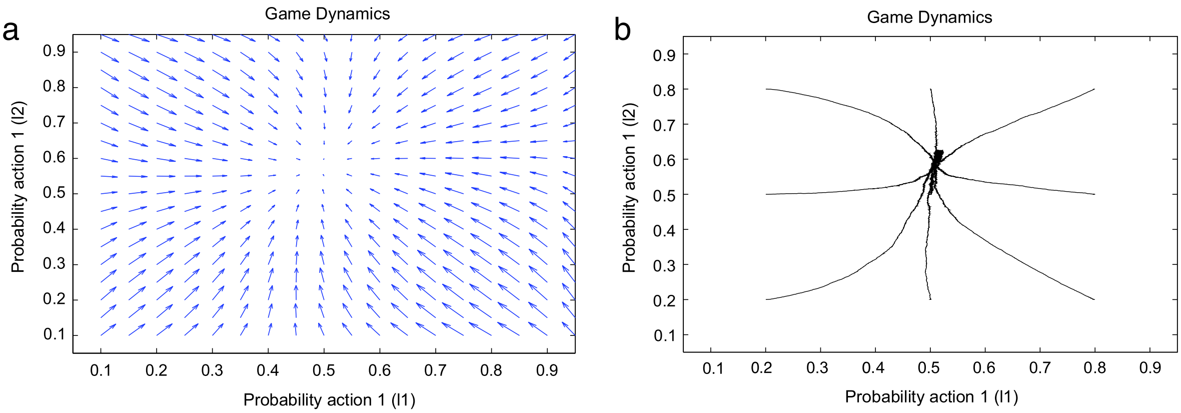

Fig. 3. (a) Predicted dynamics on the game in Table 2. (b) Experimental results on the problem in Fig. 1.

of 100 ants. The observed evolution of action probabilities on the single-colony foraging experiment closely follow these predicted dynamics.

5.2. Ant routing model

As a final demonstration we give some results for our model on a larger problem. For this experiment we model a system inspired by the AntNet routing algorithm [26], also described above. The algorithm uses ant inspired behavior to route packets in communication networks. The system builds routing tables by periodically sending out ants to explore the network. These ant packets are sent by nodes to random destinations and use the same queues as network traffic. Ants select edges in the network probabilistically based on pheromone tables which associate a pheromone value with each edge for each destination node. After arriving at a destination ants are sent backwards to update pheromone tables in visited nodes, based on the total delay for the path they followed. Data packets in the network are then routed by using a greedy policy with the pheromone tables built by ants.

We model this system using the multi-colony approach described in Section 4. Since pheromone values in AntNet depend on the destination nodes, this maps to a system with one colony for each destination in the network. Ants belonging to different colonies start from random nodes and travel through the network until they reach the destination node associated with their colony. We demonstrate this approach on the NSFNET network shown in Fig. 4(a). This former North American backbone network was one of the first networks on which AntNet was demonstrated. We limit our experiments to 2 colonies which build routing tables to route packets towards destination nodes 12 and 14. Ants from both colonies start from all other network nodes and travel to the network until they reach their destination where they receive a reward +1. When ants traverse a network edge they suffer a delay as indicated in Fig. 4(a). So in order to maximize the average reward over time, ants need to minimize the delay from each node to the destination. To simulate the influence of heavy network load slowing down the network, ants receive a penalty when both colonies use the same edge. If an ant uses an edge which more than 50% of ants from the other colony also prefer, they suffer a delay of 100 for that edge instead of the normal delay.

Even in this limited problem with only 2 colonies, explicitly calculating the approximating games becomes impractical as it would result in a game with 2 × 14 players, resulting in a large number of plays to evaluate. When we use an update like L_{R-I}, however, we can still give guarantees based on the convergence properties of the update scheme in games. We know that this update converges to a Nash equilibrium between colony policies. This means that each colony will prefer the minimum delay routes to their destination, with respect to the current policy of the other colony. A colony will thus prefer lower delay routes and avoid sharing edges since this results in high delays. Note that these results represent equilibria (i.e. local optima) but not necessarily (Pareto) optimal results for this problem. In the next section we discuss some possibilities which could be used to improve the outcome based on this equilibrium convergence behavior. A typical solution found by the L_{R-I} system is shown in Fig. 4(b). Average results over 20 runs are given in Table 5. For comparison purposes the table also lists the results obtained when both colonies use Dijkstra shortest path routing to find the minimum cost path to their destination, but do not take into account which edges are used by the other colony. From Table 5 it is clear that the pheromone based routing achieves relatively low delay routes. The antnet routing model shares an average of 0.3 edges over 20 trials, compared to 6 edges shared in shortest path routing scheme.

6. Discussion

In this paper we have shown how the behavior of pheromone based stigmergetic systems can be analyzed in terms of the dynamics of the pheromone update in an approximating game. While the explicit calculation of these games becomes cumbersome or even impossible in large problems, this analysis can still be applied by studying the dynamics of the pheromone update. For instance, as we have shown for the case of the L_{R-I} update scheme, convergence to a pure Nash

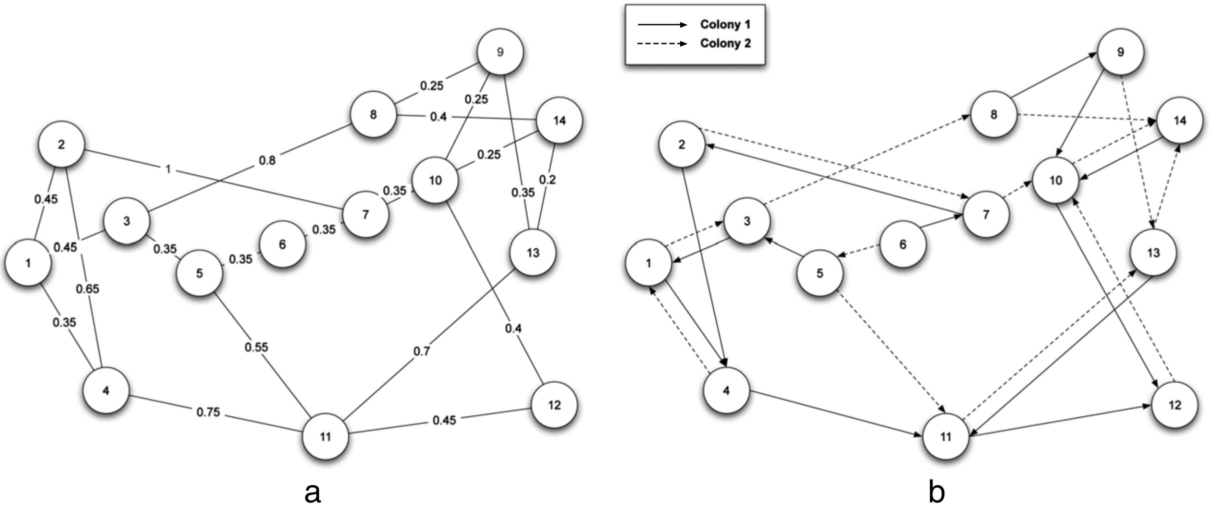

Fig. 4. (a) NFSNet former backbone (after [26]). Each link represents 2 directed edges, numbers indicate delay of an edge. (b) Example solution found by AntNet routing model using L_{R-I} update ($\lambda = 0.001$) with 2 colonies routing paths to nodes 12, 14.

Table 5
Results obtained by AntNet model in NFSNET experiment. Columns 2 and 3 give the average delay (standard deviation) to destination nodes 12 and 14, respectively (results averaged over 20 runs). For comparison purposes columns 4 and 5 give the delays that result from shortest paths based routing to both destinations, without taking into account delays caused by sharing edges.

Start node	Colony 1	Colony 2	Dijkstra 1	Dijkstra 2
1	41.2 (21.3)	41.6 (20.0)	116	33
2	44.0 (22.5)	49.7 (27.15)	208	205
3	30.8 (5.0)	27.8 (6.12)	27	24
4	38.6 (20.2)	47.3 (28.3)	109	214
5	28.1 (11.1)	31 (2.1)	20	212
6	40.1 (33.6)	34.7 (20.3)	208	205
7	43.5 (35.6)	45.6 (35.4)	108	105
8	19.3 (3.9)	9.1 (2.7)	113	8
9	15.3 (5.6)	11.3 (1.0)	108	105
10	8.0 (0.0)	5.0 (0.0)	8	5
11	11.4 (7.2)	18.4 (1.2)	9	114
12	0 (0.0)	13.0 (0.0)	0	13
13	21.8 (1.8)	4.7 (2.9)	113	100
14	13.7 (3.1)	0 (0.0)	13	0

equilibrium is a sufficient condition to assure optimal convergence in the single-colony case and convergence to a Nash equilibrium between colony policies in the multiple-colony case. These properties do not depend on the underlying game, and can therefore be assured in other problems without the need to calculate the approximating game.

In the multi-colony case one could argue that while Nash Equilibrium convergence is an interesting property, it is not necessarily the desired result for an optimization approach. In a Nash equilibrium agents play mutual best replies, and as such it represents a local optimum. Nash equilibria do not guarantee equal payoffs for the players, however, and can result in unfair solutions. Moreover, multiple equilibria can exist, with players having different preferences for these equilibria, resulting in a selection problem. Furthermore equilibria can be Pareto dominated. This means that solutions exist where all players receive at least the same reward and one or more players do strictly better (this is for instance the case in the Prisoner's Dilemma game of Table 1). Especially in the case of multi-objective optimization problems the goal is often to find a Pareto optimal (i.e. nondominated) solution.

The (pure) Nash equilibrium convergence of a learning rule, however, can be used as a basis for designing more complex systems which exhibit the desired properties. In team games, where all players receive the same payoffs and a globally optimal equilibrium always exists, noise can be added to the pheromone vectors [27]. This method, akin to simulated annealing, allows agents to escape local optima and guarantees convergence to the optimal solution even in the multi-colony case. In conflicting interest games coordination of exploration strategies can be used to find Pareto optimal solutions or periodic solutions which equalize the average payoff over time [11].

The analysis developed in this paper offers a new framework in which the outcomes of pheromone based agent coordination can be studied. Unfortunately, it is difficult to directly apply this model to existing ACO type optimization algorithms, since these systems typically incorporate additional techniques, such as heuristic information and local search. Further research is needed to determine how these extensions influence the game structure obtained for the pheromone dynamics. Biasing the agents' action selections using heuristic information could speed up the search process, provided that the heuristic correctly predicts solution rewards. However, in practice this property is difficult to guarantee, and a system using a poor heuristic may not be able to recover. While it currently is not possible to give explicit convergence guarantees for such extended systems, we believe that our framework can aid in the design of these algorithms, as it allows us to better understand the dynamics of the pheromone interactions, and to ensure that these interactions lead to desirable outcomes.

Acknowledgement

Peter Vrancx is funded by a Ph.D grant of the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT Vlaanderen).

References

[1] P. Grassé, La reconstruction du nid et les coordinations interindividuelles chez Bellicositermes natalensis et Cubitermes sp. la théorie de la stigmergie: essai d'interprétation du comportement des termites constructeurs, Insectes Sociaux 6 (1) (1959) 41–80.
[2] G. Theraulaz, E. Bonabeau, A brief history of stigmergy., Artificial Life 5 (2) (1999) 97–116.
[3] E. Bonabeau, M. Dorigo, T. Theraulaz, From Natural to Artificial Swarm Intelligence, Oxford University Press, New York, 1999.
[4] L. Panait, S. Luke, A pheromone-based utility model for collaborative foraging, in: Autonomous Agents and Multiagent Systems, 2004, AAMAS 2004, Proceedings of the Third International Joint Conference on (2004), pp. 36–43.
[5] P. Valckenears, M. Kollingbaum, Multi-agent coordination and control using stigmergy applied to manufacturing control, Mutli-Agents Systems and Applications (2001) 317–334.
[6] K. Verbeeck, A. Nowé, Colonies of learning automata, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 32 (6) (2002) 772–780.
[7] O. Holland, C. Melhuish, Stigmergy, self-organization, and sorting in collective robotics, Artificial Life 5 (2) (1999) 173–202.
[8] M. Dorigo, T. Stützle, Ant Colony Optimization, Bradford Books, 2004.
[9] M. Littman, Markov games as a framework for multi-agent reinforcement learning, in: Proceedings of the 11th International Conference on Machine Learning, 1994, pp. 322–328.
[10] K. Tuyls, Multiagent reinforcement learning: A game theoretic approach, Ph.D. Thesis, Computational Modeling Lab, Vrije Universiteit Brussel, Belgium, 2004.
[11] K. Verbeeck, A. Nowé, J. Parent, K. Tuyls, Exploring selfish reinforcement learning in repeated games with stochastic rewards, The Journal of Autonomous Agents and Multi-Agent Systems 14 (3) (2007) 239–269.
[12] P. Vrancx, K. Verbeeck, A. Nowé, Decentralized learning in markov games, IEEE Transactions on Systems, Man and Cybernetics (Part B: Cybernetics) 38 (4) (2008) 976–981.
[13] H. Gintis, Game Theory Evolving: A Problem-Centered Introduction to Modeling Strategic Behavior, Princeton University Press, Princeton, New Jersey, 2000.
[14] J. Osborne, A. Rubinstein, A Course in Game Theory, MIT Press, Cambridge, MA, 1994.
[15] J. Nash, Equilibrium points in n-person games, Proceedings of the National Academy of Siences 36 (1950) 48–49.
[16] M. Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming, John Wiley & Sons, Inc., New York, NY, USA, 1994.
[17] J. Filar, D. Krass, Hamiltonian cycles and markov-chains, Mathematics of Operations Research 19 (1) (1994) 223–237.
[18] K. Narendra, M. Thathachar, Learning Automata: An Introduction, Prentice-Hall International, Inc, 1989.
[19] M. Dorigo, V. Maniezzo, A. Colorni, Ant system: optimization by a colony of cooperating agents, IEEE Transactions on Systems, Man, and Cybernetics, Part B 26 (1) (1996) 29–41.
[20] A. Nowé, K. Verbeeck, Formalizing the ant algorithms in term of reinforcement learning, in: Proceedings of the 5th European Conference on Artificial life, Springer-Verlag LNAI1674, Lausanne, Switzerland, 1999, pp. 616–620.
[21] M. Thathachar, P. Sastry, Networks of Learning Automata: Techniques for Online Stochastic Optimization, Kluwer Academic Publishers, 2004.
[22] P. Sastry, V. Phansalkar, M. Thathachar, Decentralized learning of nash equilibria in multi-person stochastic games with incomplete information, IEEE Transactions on Systems, Man, and Cybernetics 24(5) (1994) 769–777.
[23] R. Wheeler, K. Narendra, Decentralized learning in finite markov chains, IEEE Transactions on Automatic Control AC-31 (1986) 519–526.
[24] A. Schwartz, A reinforcement learning method for maximizing undiscounted rewards, in: Proceedings of the Tenth International Conference on Machine Learning, 1993, pp. 298–305.
[25] D. Angus, C. Woodward, Multiple objective ant colony optimisation, Swarm Intelligence 3 (1) (2009) 69–85.
[26] G.D. Caro, M. Dorigo, Antnet: Distributed stigmergetic control for communications networks, Journal of Artificial Intelligence Research 9 (2) (1998) 317–365.
[27] M. Thathachar, V. Phansalkar, Learning the global maximum with parameterized learning automata, IEEE Transactions on Neural Networks 6 (2) (1995) 398–406.