Reinforcement Social Learning of Coordination in Networked Cooperative Multiagent Systems

Jianye Hao[1], **Dongping Huang**[2], **Yi Cai**[2], **Ho-fung Leung**[3]
[1]Massachusetts Institute of Technology
jianye@mit.edu
[2]South China University of Technology
{*huang.dp, ycai*}*@mail.scut.edu.cn*
[3]The Chinese University of Hong Kong
lhf@cuhk.edu.hk

Abstract

The problem of coordination in cooperative multiagent systems has been widely studied in the literature. In practical complex environments, the interactions among agents are usually regulated by their underlying network topology, which, however, has not been taken into consideration in previous work. To this end, we firstly investigate the multiagent coordination problems in cooperative environments under the *networked social learning* framework focusing on two representative topologies: the small-world and the scale-free network. We consider a population of agents where each agent interacts with another agent randomly chosen from its neighborhood in each round. Each agent learns its policy through repeated interactions with its neighbors via social learning. It is not clear a priori if all agents can learn a consistent optimal coordination policy and what kind of impact different topology parameters could have on the learning performance of agents. We distinguish two types of learners: *individual action learner* and *joint action learner*. The learning performances of both learners are evaluated extensively in different cooperative games, and the influence of different factors on the learning performance of agents is investigated and analyzed as well.

1 Introduction

In multiagent systems (MASs), one key property of an agent is to be able to adaptively adjust its behaviors according to others' behaviors to achieve effective coordination on desirable outcomes. One central and widely studied class of coordination problem is how to coordinate within cooperative MASs, in which the agents share common interests (Claus and Boutilier 1998; Matignon, Laurent, and For-Piat 2012). In cooperative MASs, the agents share common interests and the same reward function, and the increase in individual satisfaction also results in the increase in the satisfaction of the group.

Until now, various multiagent reinforcement learning algorithms (Claus and Boutilier 1998; Lauer and Riedmiller 2000; Kapetanakis and Kudenko 2002; Wang and Sandholm 2002; Brafman and Tennenholtz 2004; Matignon, Laurent, and For-Piat 2012) have been proposed to tackle the coordination problem in cooperative MASs. The most commonly adopted learning framework for studying the coordination problem within cooperative MASs is to consider two (or more) players playing a repeated (stochastic) game, in which the agents learn their optimal coordination policies through repeated interactions with the same opponent(s) (Matignon, Laurent, and For-Piat 2012). However, in practical complex environments, it is unlikely for an agent to always interact with the same partner, and the interacting partners of different agents may vary frequently. The non-fixed partner interaction adds additional complexity to the coordination in cooperative MASs, since an agent's policy that achieves coordination on an optimal joint action with one partner may fail when it comes to a different partner. Hao and Leung [2013] make the first step in proposing the social learning framework to investigate the multiagent coordination problem in cooperative MASs in which each agent learns its policy through repeated interactions with randomly chosen partners. However, no underlying interaction topology is considered in their social learning framework, which thus cannot fully reflect the interaction in practical multiagent systems (Olfati-Saber, Fax, and Murray 2007). To make the coordination techniques applicable in practice especially for those MAS applications closely residing on existing social networks, it is important to explicitly take into consideration the underlying topology of interaction environment when desinging coordination techniques. Until now it is still not clear a priori if and how the agents are able to coordinate on and converge to optimal solutions under different topologies, since different topologies may have predominant impact on the coordination performance among agents. Another important question is what kind of impact that different topologies and topology parameters could have on the learning performance of agents in different cooperative environments.

To this end, in this paper, we firstly study the multiagent coordination problem in cooperative games within the *networked social learning framework* by taking the underlying topology into consideration. In each round each agent interacts with one of its neighbors randomly, and the interactions between each pair of agents are modeled as two-player cooperative games. Each agent learns its policy concurrently over repeated interactions with randomly selected neighbors. Under the networked social learning framework, each agent usually only has the opportunity to interact with a very small proportion of agents and also different agents interact with different proportions of agents depending on their own connection degrees. Besides, each agent may also learn from the experience of its neighbors. We distinguish two different types of learning environments within the *networked social learning framework* depending on the amount of information the agents can perceive, and propose two types of learners accordingly: individual action learners (IALs) and joint action learners (JALs). IALs learn the values of each individual action directly by considering their neighbors as part of the environment, while JALs learn the values of each action indirectly based on the values of the joint actions. Both IALs and JALs employ the optimistic assumption and the FMQ heuristic to utilize the learning experience of their own and their neighbors. Two representative social networks models are considered: the small-world and scale-free network. We extensively evaluate the learning performances of both types of learners in different types of cooperative games within the networked social learning framework for both topologies. The experimental results and analysis also shed light on the impact of different factors (i.e., the underlying topology, different parameters, and IAL/JAL) on the learning dynamics of agents in different cooperative games.

The remainder of the paper is organized as follows. In Section 2, we give an overview of previous work of coordination in cooperative MASs. In Section 3, the networked social learning framework and both IALs and JALs are described. In Section 4, we present the evaluation results of both types of learners in different cooperative games and investigate the influence of different factors. Lastly conclusion and future work are given in Section 5.

2 Related Work

Until now significant research efforts have been devoted to solve the coordination problem in cooperative MASs in the multiagent learning literature. Usually the cooperative multiagent environment is modeled as two-player cooperative repeated (or stochastic) games. Claus and Boutilier (1998) firstly distinguished two different types of learners (without optimistic exploration) based on Q-learning algorithm: independent learner and joint-action learner, and investigate their performance in the context of two-agent repeated cooperative games. Empirical results show that both learners can successfully coordinate on the optimal joint actions in simple 2×2 cooperative games. However, both of them fail to coordinate on optimal joint actions when the game structure becomes more complex i.e., the climbing game and the penalty game. Following that, a number of improved learning algorithms have been proposed. Lauer and Riedmiller [2000] proposed the distributed Q-learning algorithm based on the optimistic assumption where each action's Q-value is updated in such a way that only the maximum payoff received by performing this action is considered. Besides, an additional coordination mechanism is required for agents to avoid mis-coordination on suboptimal joint actions. The authors proved that it is guaranteed to coordinate on optimal joint actions if the cooperative game is deterministic. However, it fails when dealing with stochastic environments.

Later a number of approaches (Kapetanakis and Kudenko 2002; Matignon, Laurent, and For-Piat 2008; Panait, Sullivan, and Luke 2006) have been proposed to handle the stochasticity of the games. One representative work under this direction was that of Kapetanakis and Kudenko (2002), who propose the FMQ heuristic to alter the Q-value estimation function to handle the stochasticity of the games. Under the FMQ heuristic, the original Q-value for each individual action is modified by incorporating the additional information of how frequent the action receives its corresponding maximum payoff. Experimental results show that FMQ agents can successfully coordinate on an optimal joint action in partially stochastic climbing games, but fail in fully stochastic climbing games. Matignon et al. (2012) review all existing independent multiagent reinforcement learning algorithms in cooperative MASs, and evaluate and discuss their strength and weakness. Their evaluation results show that all of them fail to achieve coordination for fully stochastic games and only recursive FMQ can achieve coordination for 58% of the runs.

All the above previous works only focus on the case of coordinating towards optimal joint actions in cooperative games under the *(two) fixed-agent repeated learning framework*. Most recently Hao and Leung (2013) have firstly proposed and investigated the problem of coordinating towards optimal joint actions in cooperative games within the *social learning framework*. However, in their framework, the agents' interactions are purely random, without considering any underlying interaction topology, which thus fail to accuratley reflect the interaction scenarios in practical MASs (Olfati-Saber, Fax, and Murray 2007). It is not clear a priori how the agents are able to converge to optimal solutions and how the learning performance would be affected in different cooperative games when different underlying topologies of the interaction environment are considered.

3 Networked Social Learning Framework

Under the networked social learning framework, there are a population of N agents in which each agent's neighborhood is determined by the underlying network topology. Each agent learns its policy through repeated pairwise interactions with its neighbors in the population. The interaction between each pair of agents is modeled as a two-player cooperative game. Each agent i knows its own action set A_i, but have no access to their payoffs under each outcome beforehand. During each round, each agent interacts with another agent randomly chosen from its neighbors, and one agent is randomly assigned as the row player and the other agent as the column player. The agents are assumed to know their roles, i.e., either as row player or column player, during each interaction. Without loss of generality, we assume that both the row and column agents share the same set of actions in the cooperative game being played. At the end of each round, each agent updates its policy based on the learning experience it receives from the current round. The overall interaction protocol under the networked social learning framework is presented in Algorithm 1.

As previously mentioned, studies show that most of the real-world networks are not either regular (e.g., lattice network) or purely random (Wang and Chen 2003), which lead us to think that regular or random topologies are oversimplified and thus are not the most ideal candidates for modeling practical interactions in MASs. Therefore, in this work, we focus on two realistic network topologies: small-world networks (SWN) and scale-free networks (SFN), which have been shown to reflect a large number of real-world networks (Albert and Barabási 2002). Besides, compared with the traditional fixed-agent repeated interaction framework, under the networked social learning framework, the agents may also be able to learn from their neighbors. Therefore, in Section 3.2, we distinguish two different learning settings in terms of the amount of information each agent can perceive. Following that, we propose two classes of learners, *individual action learners* (IALs) and *joint action learners* (JALs) in Section 3.3, which are applicable for the two learning settings we distinguish respectively.

Algorithm 1 Overall interaction protocol of the networked social learning framework

```
for a number of rounds do
    for each agent in the population do
        One neighbor is randomly chosen as its interacting partner, and one of them is assigned as the row player and the other one as the column player.
        Both agents play a two-player cooperative game by choosing their actions from their own action set independently and simultaneously.
    end for
    for each agent in the population do
        Update its policy based on its experience in the current round
    end for
end for
```

3.1 Interaction Networks

The interaction networks determine the possible interactions among different agents and also the amount of information each agent can perceive in the system. Different interaction topologies may have significant influence on the collective learning performance of agents in a cooperative multiagent system. We consider the following two complex network models: small-world network and scale-free network, which are able to accurately capture the major properties of a large variety of real-world networks (Albert and Barabási 2002).

Small-world Networks It reflects the "what a small world" phenomenon reflected in many practical networks including collaboration networks (e.g., the co-authorship of research papers) and the social influence networks (Albert and Barabási 2002). This kind of networks are featured by high clustering coefficients and short average path lengths. Another characteristic of a small-world network is that its connectivity (degree) distribution peaks at an average value and decays exponentially on both sides, i.e., most of the nodes have the same number of connections. A small-world network can usually be represented as $SW_N^{k,\rho}$, where N is the size of the network, k is its average connectivity degree and ρ is the re-wiring probability indicating the degree of the network randomness.

Scale-free Networks Different from small-world networks, the connectivity distribution of a scale-free network follows the power law distribution, i.e., most of the nodes have very few connections while only a few nodes have very large connections. This kind of "scale-free" feature has been observed in many real-world networks such as the connection network of web pages (Barabási, Albert, and Jeong 2000) and citation network of research papers (Redner 1998). For each node, the probability of being connected to k adjacent nodes is proportional to $k^{-\gamma}$, and we denote a scale-free network as SF_N^{γ}, where N is the network size.

3.2 Observation Mechanism

Under the networked social learning framework, each agent interacts with one of its neighbors randomly chosen during each round. We define each pair of interacting agents as being in the same group. Within a social learning environment, since every agent interacts with its own interaction partner simultaneously, different agents may be exposed to interaction experience of agents from other groups through communications or observations (Villatoro, Sabater-Mir, and Sen 2011; Hao and Leung 2013). Allowing agents to observe the information of other agents outside their direct interactions may result in a faster learning rate and facilitate coordination on optimal solutions. In the *networked social learning framework*, it is thus reasonable to assume that each agent can have access to its neighbors' learning experience. Notice that the amount of learning experience available to each agent varies and depends on its neighborhood size.

We identify two different learning settings depending on the amount of information that each agent can perceive. In the first setting, apart from its own action and payoff, each agent can also observe the actions and payoffs of all neighbors with the same role as itself. Formally, the information that each agent i can perceive at the end of each round t can be represented as the set $S_i^t = \{\langle s_i^t, a_i^t, r^t\rangle, \langle s_{b,1}^t, b_1^t, r_1^t\rangle, \ldots, \langle s_{b,N(i)}^t, b_{N(i)}^t, r_{N(i)}^t\rangle\}$. Here $\langle s_i^t, a_i^t, r^t\rangle$ are agent i's current state, its action and payoff, and the rest are the corresponding current states, actions and payoffs of all its neighbors. This setting is parallel to the individual action learning setting under the fixed agent repeated interaction framework (Claus and Boutilier 1998) and the social learning framework without any underlying topology (Hao and Leung 2013), and can be considered as a natural extension to the networked social learning environment based on the observation mechanism.

The second setting is a natural extension of the joint action learning setting in both the fixed agent repeated interaction framework (Claus and Boutilier 1998) and the social learning framework without any underlying topology (Hao and Leung 2013) to the networked social learning framework. Apart from the same information available in the first setting, each agent is also assumed to able to perceive the action of its interaction partner and those agents with opposite role from its neighbors' groups. Formally, the experience for each agent i at the end of each round t can be denoted as the set $P_i^t = \{\langle s_i^t, (a_i^t, a_j^t), r^t\rangle, \langle s_{b,1}^t, (b_1^t, c_1^t), r_1^t\rangle, \ldots,$ $\langle s_{b,N(i)}^t, (b_{N(i)}^t, c_{N(i)}^t), r_{N(i)}^t\rangle\}$. Here $\langle s_i^t, (a_i^t, a_j^t), r^t\rangle$ consists of the current state s_i^t of agent i, the joint action of agent i and its partner j, and agent i's payoff. The rest consists of the current state, the joint actions and payoffs of all its neighbors' groups respectively.

3.3 Learning Strategy

In general, to achieve coordination on optimal joint actions, an agent's behaviors as the row or column player may be the same or different, depending on the characteristics of the game and its opponent' behavior. Accordingly we propose that each agent should employ a pair of strategies, one used when the agent is the row player and the other used when it is the column player, to play with any other agent in its neighborhood. The strategies we develop here are natural extensions of the Q-learning techniques (Watkins and Dayan 1992) to the networked social learning framework. There are two distinct ways of applying Q-learning techniques to the networked social learning framework depending on the learning setting that the agents are situated in as we described in Section 3.2.

Individual Action Learner In the first setting, each agent only perceives the actions and payoffs of itself and its neighbors. Thus it is reasonable for each agent to simply consider its interaction partner as part of the environment. Each agent has two possible states corresponding to its roles as the row player or column player, and each agent knows its current role during each interaction. Naturally each agent holds a Q-value $Q(s, a)$ for each action a under each state $s \in \{Row, Column\}$, which keeps record of action a's past performance and serves as the basis for making decisions. At the end of each round t, each agent i picks action a^* (randomly choosing one action in case of a tie) with the highest payoff among all its neighbors with the same role as itself from S_i^t, and updates this action's Q-value under its current state s_i^t using Equation 1,

$$Q_i^{t+1}(s_i^t, a^*) = Q_i^t(s_i^t, a^*) + \alpha_i^t(s_i^t) \times [r_{max}^t(s_i^t, a^*) \times f_i^t(a^*) - Q_i^t(s_i^t, a^*)] \quad (1)$$

where

- $r_{max}^t(s_i^t, a^*) = \max\{r \mid \langle s', a^*, r\rangle \in S_i^t, s' = s_i^t\}$, which is the highest payoff received by choosing action a^* under the same role of s_i^t based on the set S_i^t,
- $f_i^t(a^*) = \frac{|\{\langle s',a^*,r\rangle | \langle s',a^*,r\rangle \in S_i^t, s'=s_i^t, r=r_{max}^t(s_i^t,a^*)\}|}{|\{\langle s',a^*,r\rangle | \langle s',a^*,r\rangle \in S_i^t, s'=s_i^t\}|}$, which is the empirical frequency of receiving the reward of $r_{max}^t(s_i^t, a^*)$ by choosing action a^* based on the current round experience,
- $\alpha_i^t(s_i^t)$ is agent i's current learning rate in state s_i^t.

The above Q-value update rule intuitively incorporates both the optimistic assumption and the FMQ heuristic (Kapetanakis and Kudenko 2002). On one hand, this update rule is optimistic since we only update the Q-value of the action that receives the highest payoff based on the current round's experience, and also its Q-value is updated based on the highest payoff only. On the other hand, similar to the FMQ heuristic, the update rule also takes into account the information of how frequent the corresponding highest payoff can be received.

Each agent chooses its action based on the set of Q-values corresponding to its roles during each interaction according to the ϵ-greedy mechanism. Specifically with probability $1 - \epsilon$ each agent chooses its action with the highest Q-value to exploit the action with best performance currently (tie is broken randomly), and with probability ϵ makes random choices for the purpose of exploring new actions with potentially better performance.

Joint Action Learner In the joint action learning setting, each agent has more information at its disposal since it can have access to the joint actions of its own group and its neighbors' groups. Consequently, each agent can learn the Q-values for each joint action in contrast to learning Q-values for individual actions only in the individual action learning setting. Specifically, at the end of each round t, each agent i updates its Q-values under its current state s_i^t for each joint action $\overrightarrow{a}$ which satisfies the constraint of $\langle s_i^t, \overrightarrow{a}, r(\overrightarrow{a})\rangle \in P_i^t$ as follows,

$$Q_i^{t+1}(s_i^t, \overrightarrow{a}) = Q_i^t(s_i^t, \overrightarrow{a}) + \alpha_i^t(s_i^t) \times [r(\overrightarrow{a}) - Q_i^t(s_i^t, \overrightarrow{a})] \quad (2)$$

where $r(\overrightarrow{a})$ is the payoff of agent i (or a neighbor if the information is obtained through obversation mechanism) under state s under the joint action $\overrightarrow{a}$ and $\alpha_i^t(s_i^t)$ is its current learning rate under state s_i^t.

After enough explorations, it is expected that the above Q-values can reflect the expected performance of each joint action, but each agent still needs to determine the relative performance of each individual action to make informed decisions. At the end of each round t, for each action a, define $r_a^{max}(s_i^t) = max\{Q_i^{t+1}(s, (a, b)) \mid b \in A_i\}$, and denote the corresponding opponent's action as $b^{max}(a)$. The value of $r_a^{max}(s)$ reflects the maximum possible expected payoff that agent i can obtain by performing action a under the current state s_i^t. However, agent i's actual expected payoff of performing action a generally depends on the action choices of its interacting partners. To take this factor into consideration, each agent i also maintains the belief of the frequency of its interacting partners performing action b when it chooses action a, which is denoted as $f_i(s_i^t, \langle a, b\rangle)$. The value of $f_i(s_i^t, \langle a, b\rangle)$ is estimated based on agent i's current round experience P_i^t as follows,

$$f_i(s_i^t, \langle a, b\rangle) = \frac{|\{\langle s', (a, b), r\rangle \mid \langle s', (a, b), r\rangle \in P_i^t, s' = s_i^t\}|}{|\{\langle s', (a, y), r\rangle \mid \langle s', (a, y), r\rangle \in P_i^t, y \in A_i, s' = s_i^t\}|} \quad (3)$$

Finally, each agent i assesses the relative performance $EV(s, a)$ of an action a under the current state s_i^t as follows,

$$EV(s_i^t, a) = r_a^{max}(s_i^t) \times f_i(s_i^t, \langle a, b^{max}(a)\rangle) \quad (4)$$

Overall JALs evaluate the relative performance of each action based on both the optimistic assumption and the information of the frequency that the maximum payoff can be received by performing this action. Based on the EV-values of each individual action, each agent chooses its action in a similar way as it would use Q-values for IALs following the ϵ-greedy mechanism.

4 Experimental Results

In this section, we present the evaluation results of IALs and JALs in different types of cooperative games and also inves-

1's payoff / 2's payoff		Agent 2 a	b	c
Agent 1	a	11	-30	0
	b	-30	7	6
	c	0	0	5

(a) CG

1's payoff / 2's payoff		Agent 2 a	b	c
Agent 1	a	10	0	k
	b	0	2	0
	c	k	0	10

(b) PG

1's payoff / 2's payoff		Agent 2 a	b	c
Agent 1	a	11	-30	0
	b	-30	14/0	6
	c	0	0	5

(c) PSCG

1's payoff / 2's payoff		Agent 2 a	b	c
Agent 1	a	10/12	5/-65	8/-8
	b	5/-65	14/0	12/0
	c	5/-5	5/-5	10/0

(d) FSCG

Fig. 1: Different Types of Cooperative Games ((c) (b, b) yields the payoff of 14 or 0 with equal probability (d) Each outcome yields two different payoffs with equal probability)

tigate the influences of different parameters under the networked social learning framework. Unless otherwise mentioned, for the small-word network $SW_N^{k,\rho}$, the default setting is $N = 200, k = 6, \rho = 0.2$, and for the scale-free network SF_N^{γ}, the default setting is $N = 200, \gamma = 3$. For all agents, the initial learning rate α is 0.9 and the exploration rate ϵ is initially set to 0.4. Both values are exponentially decreased until 0. The initial Q-values are randomly generated between -10 and 10. All results are averaged over 200 times.

4.1 Performance Evaluation

Deterministic Games We first consider two representative and particularly difficult deterministic coordination problems: the climbing game (CG) (Fig. 1a) and the penalty game (PG) with $k = -50$ (Fig. 1b). The climbing game has one optimal joint action (a, a) and two joint actions (a, b) and (b, a) with high penalties. The high penalty induced by (a, b) or (b, a) can make the agents find action a very unattractive, which thus may result in convergence to the suboptimal outcome (b, b). Fig. 2a and 2b show the average payoffs of IALs and JALs as functions of the number of rounds under both small-world and scale-free networks for the climbing game and penalty game respectively. We can see that both IALs and JALs can receive the highest average payoff of 11 (successfully coordinating on the optimal joint action (a, a)) under both networks. Besides, JALs learn faster towards optimal outcomes than IALs in both networks since JALs have more information at their disposal.

Stochastic Games Next we consider two stochastic variants of the climbing game - the partially stochastic climbing game (PSCG) (Fig. 1c) and fully stochastic climbing game (FSCG) (Fig. 1d). Both games are in essence equivalent to the original climbing game, since the expected payoff of each agent under each outcome remains unchanged. However, it is much more difficult for the agents to converge to the optimal outcome due to the stochastic feature introduced. For example, in Fig. 1c the joint action (b, b) yields the payoff of 14 with probability of 0.5, which makes it easy for the agents to misperceive (b, b) as the optimal joint action.

Fig. 2c and 2d illustrate the average payoffs of both IALs and JALs as functions of the number of rounds in both networks for the partially and fully stochastic climbing games respectively. First we can see that both IALs and JALs can achieve the highest average payoff of 11 (reach full coordination on (a, a)) for both stochastic games in both networks. Another observation is that the JALs do perform significantly better than that of IALs in terms of the convergence rate. This is expected since the JALs can distinguish the Q-values of different joint actions and have the ability of quickly identifying which action pair is optimal. In contrast, for the IALs, since they cannot perceive the actions of their interacting partners, it is more difficult for them to distinguish between the noise from the stochasticity of the game and the explorations of their interacting partners. Thus it takes more time for the IALs to learn the actual Q-values of their individual actions.

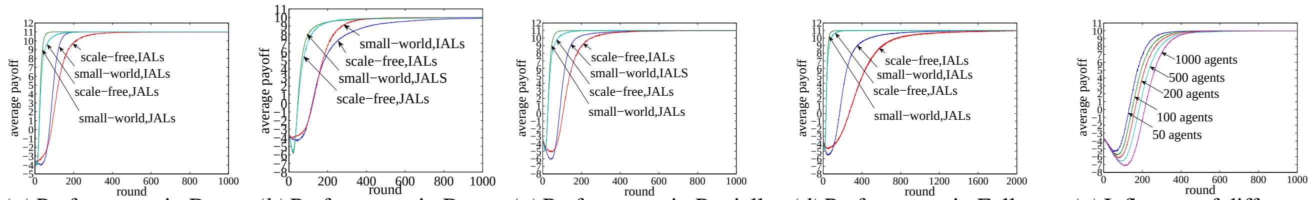
Summary For both deterministic and stochastic games, one interesting observation is that both IAL and JALs usually learn faster in small-world network. We hypothesize that it is due to the fact that in our setting the average number of neighbors in the small-world network is more than that in the scale-free network. According to our observation mechanism, each round the agents in small-world network thus have more experience to learn from and thus can learn faster towards optimal outcomes. Finally, it is worth mentioning that for the same game, the learning strategy proposed in (Hao and Leung 2013) based on individual actions where the topology is not considered always fails to converge to the optimal joint action (a, a). This confirms that the network topology has significant influence on the agents' learning performance, and our networked social learning framework actually facilitates better coordination among agents than theirs.

4.2 Influences of Different Parameters

In this section, we turn to investigate the influence of different topology parameters on the learning performance of agents under the networked social learning framework. Due to space limitation, we only present the results for IALs and JALs in penalty game in the small-world network, but the general conclusions are similar for the scale-free network.

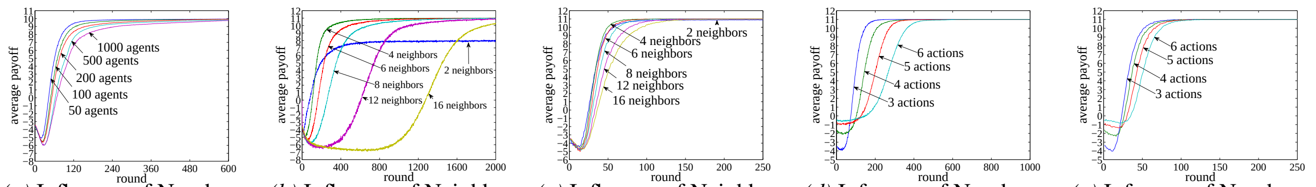
Influences of the size of the population Fig. 2e and 3a show the dynamics of IALs' and JALs' average payoffs with different population sizes in the small-world network respectively. We can easily observe the trend that the convergence rate is decreased with the increase of the total number of agents. This is reasonable since both IAL and JAL learn based on their local information only, and the larger the population size becomes, the more difficult it is for them to coordinate the actions of all agents in the population towards a consistent optimal solution.

Influences of the neighborhood size Fig. 3b and 3c show the dynamics of IALs' and JALs' average payoffs when the neighborhood size varies in the small-world network respectively. It is interesting to observe that both IALs and JALs' learning performance in terms of converging to optimal outcome is initially increased with the increase of the neighborhood size ($2 \rightarrow 4$), but gradually decreased when the neighborhood size is further increased ($4 \rightarrow 6 \rightarrow 8 \rightarrow 12 \rightarrow 16$).



(a) Performance in Deterministic Climbing Game *(b)* Performance in Deterministic Penalty Game *(c)* Performance in Partially Stochastic Climbing Game *(d)* Performance in Fully Stochastic Climbing Game *(e)* Influence of different number of agents for IALs

Fig. 2: Average payoff of IALs and JALs in different games and different networks



(a) Influence of Number of Agents for JALs *(b)* Influence of Neighborhood Size for IALs *(c)* Influence of Neighborhood Size for JALs *(d)* Infuence of Number of Actions for IALs *(e)* Infuence of Number of Actions for JALs

Fig. 3: Influence of different parameters for IALs and JALs in the small-world network

The neighborhood size can be considered as the information sharing degree among agents, and too much or less information seem to both detriment the learning performance of both learners. We hypothesize that setting the neighborhood size too small or large can either result in the agents underestimating or overestimating the relative performance of different actions respectively. One useful insight of this observation is that in practical distributed systems, agents may only require few local communication with their neighbors to achieve the globally optimal coordination, given that the interactions among agents are carefully regulated by the underlying topology of the system. Another observation is that the neighborhood size has much more influence on the learning performance of IALs than that of JALs. This is reasonable because IALs are more susceptible to the noise from the stochasticity of the game and the explorations of their interacting partners, and a small change of the neighborhood size can more significantly influence their estimations of the relative values of each action compared with JALs.

Table 1: General Learning Performance Evaluation

Success Rate	DCGs in SWN	SCGs in SWN	DCGs in SFN	SCGs in SFN
IALs	100% (153 round)	70.67% (319 round)	100% (235 round)	68%(455 round)
JALs	100% (41 round)	93,67%(48 round)	100%(43 round)	95%(49 round)

Influences of the number of actions Fig. 3d and 3e show the dynamics of IALs' and JALs' average payoffs when different number of actions varies (from 3 to 6) respectively. As we can see, the convergence rate is significantly decreased with the increase of the number of actions. Intuitively the larger the action space of the learner is, the more suboptimal (yet misleading) outcomes the games may have. Therefore it may take more time and experience for the learners to learn towards a consistent optimal action pair.

4.3 Performance in General Cooperative Games

In this section, we further evaluate the performance of IALs and JALs in both networks for general deterministic and stochastic cooperative games (DCGs and SCGs). Both DCGs and SCGs are generated randomly with the payoffs ranging between -20 and 20. The success rate and convergence rate (i.e., the average number of rounds before convergence) are listed in Table 1 averaged over 100 randomly generated games, which show that both IALs and JALs can achieve full coordination on optimal solutions for all DCGs, while fail for certain percentage of SCGs. However, JALs perform much better than IALs in SCGs. Besides, the JALs' convergence rate towards optimal solutions is much higher than IALs for both DCGs and SCGs. Intuitively, the superior performance of JALs can be briefly explained as follows. In SCGs, JALs can have accurate estimation of the value of each joint action, while IALs cannot. Accordingly JALs can successfully distinguish between the environment noise (stochasticity) and the partners' explorations in SCGs, while IALs cannot. Thus it is much easier for IALs to fail in SCGs than JALs. For JALs, the small percentage of failure is due to that in some games there exist multiple outcomes very close to the optimal one, and the inaccurate estimation of the partners' behaviors may result in the miscoordination.

5 Conclusion

We are the first to investigate how agents can achieve efficient coordination on optimal outcomes in different cooperative environments under the *networked social learning framework* by proposing two types of learners: IALs and JALs. We also show that different network topology factors can have significant influence on the learning performance of agents. Compared with previous work, our framework could be useful in faciliating the coordination among agents in real-world networked MASs, and provide insights of how to design cooperative MASs towards efficient coordination.

References

Albert, R., and Barabási, A.-L. 2002. Statistical mechanics of complex networks. *Rev. Modern Phys.* 74(1):47–97.

Barabási, A.-L.; Albert, R.; and Jeong, H. 2000. Scale-free characteristics of random networks: the topology of the world-wide web. *Physica A: Statistical Mechanics and its Applications* 281(1):69–77.

Brafman, R. I., and Tennenholtz, M. 2004. Efficient learning equilibrium. *Artificial Intelligence* 159:27–47.

Claus, C., and Boutilier, C. 1998. The dynamics of reinforcement learning in cooperative multiagent systems. In *AAAI'98*, 746–752.

Hao, J. Y., and Leung, H. F. 2013. The dynamics of reinforcement social learning in cooperative multiagent systems. In *IJCAI'13*, 184–190. AAAI Press.

Kapetanakis, S., and Kudenko, D. 2002. Reinforcement learning of coordination in cooperative multiagent systems. In *AAAI'02*, 326–331.

Lauer, M., and Riedmiller, M. 2000. An algorithm for distributed reinforcement learning in cooperative multi-agent systems. In *ICML'00*, 535–542.

Matignon, L.; Laurent, G. J.; and For-Piat, N. L. 2008. A study of fmq heuristic in cooperative multi-agent games. In *AAMAS'08 workshop: MSDM*, 77–91.

Matignon, L.; Laurent, G. J.; and For-Piat, N. L. 2012. Independent reinforcement learners in cooperative markov games: a survey regarding coordination problems. *Knowledge Engineering Review* 27:1–31.

Olfati-Saber, R.; Fax, J. A.; and Murray, R. M. 2007. Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE* 95(1):215–233.

Panait, L.; Sullivan, K.; and Luke, S. 2006. Lenient learners in cooperative multiagent systems. In *AAMAS'06*, 801–803.

Redner, S. 1998. How popular is your paper an empirical study of the citation distribution. *Eur. Hpys. J. B.* 4(2):132–134.

Villatoro, D.; Sabater-Mir, J.; and Sen, S. 2011. Social instruments for robust convention emergence. In *IJCAI'11*, 420–425.

Wang, X. F., and Chen, G. R. 2003. Complex networks: small-world, scale-free and beyond. *Circuits and Systems Magazine* 3(1):6–20.

Wang, X., and Sandholm, T. 2002. Reinforcement learning to play an optimal nash equilibrium in team markov games. In *NIPS'02*, 1571–1578.

Watkins, C. J. C. H., and Dayan, P. D. 1992. Q-learning. *Machine Learning* 279–292.