2011 Workshops of International Conference on Advanced Information Networking and Applications

Achieving Context Awareness and Intelligence in Distributed Cognitive Radio Networks: A Payoff Propagation Approach

Kok-Lim Alvin Yau[1], Peter Komisarczuk[2,1] and Paul D. Teal[1]
[1]School of Engineering and Computer Science, Victoria University of Wellington, New Zealand
[2]School of Computing and Technology, Thames Valley University, UK
kok-lim.yau@ecs.vuw.ac.nz, peter.komisarczuk@tvu.ac.uk, paul.teal@ecs.vuw.ac.nz

***Abstract*-Cognitive Radio (CR) is a next-generation wireless communication system that exploits underutilized licensed spectrum to optimize the utilization of the overall radio spectrum. A Distributed Cognitive Radio Network (DCRN) is a distributed wireless network established by a number of CR hosts in the absence of fixed network infrastructure. Context awareness and intelligence are key characteristics of CR networks that enable the CR hosts to be aware of their operating environment in order to make an optimal joint action. This research aims to achieve context awareness and intelligence in DCRN using our novel Locally-Confined Payoff Propagation (LCPP), which is an important feature in Multi-Agent Reinforcement Learning (MARL). The LCPP mechanism is suitable to be applied in most applications in DCRN that require context awareness and intelligence such as scheduling, congestion control, as well as Dynamic Channel Selection (DCS), which is the focus of this paper. Simulation results show that the LCPP mechanism is a promising approach. The LCPP mechanism converges to an optimal joint action including networks with cyclic topology. Fast convergence is possible. The investigation in this paper serve as an important foundation for future work in this research field.**

***Keyword*-Cognitive radio networks; multi-agent reinforcement learning; payoff propagation**

I. Introduction

Studies sponsored by the Federal Communications Commission (FCC) pointed out that the current static spectrum allocation has led to overall low spectrum utilization where up to 70% of the licensed spectrum remains unused (called white space) at any one time even in a crowded area [1]. The white space can be defined by time and frequency at a particular location. Cognitive Radio (CR) is a next-generation wireless communication system that enables unlicensed spectrum users or Secondary Users (SUs) to use the white space of licensed users' or Primary Users' (PUs') spectrum conditional on the interference to the PU being below an acceptable level. A Distributed Cognitive Radio Network (DCRN) is a distributed wireless network comprised of a number of SUs that interact with each other in a common operating environment in the absence of fixed network infrastructure such as a base station.

Context awareness and intelligence, as achieved by the popular conceptual cognition cycle [2], are key characteristics of CR networks. Through context awareness, an SU is aware of its operating environment; and through intelligence, an SU utilizes the sensed and inferred *high* quality white space in an efficient manner without following a static pre-defined policy. The key terms in this paper are: 1) *joint action*; and 2) *optimal joint action.* A joint action is defined as the actions taken by all the SUs throughout the DCRN. An optimal joint action is the joint action that provides ideal and optimal network-wide performance. The purpose of context awareness and intelligence is to achieve an optimal joint action in a distributed manner. In this paper, our objective is to investigate our novel Locally-Confined Payoff Propagation (LCPP), which is an important component in Multi-Agent Reinforcement Learning (MARL) [3], and it is applied to design the Dynamic Channel Selection (DCS) scheme in DCRN. In [3], the MARL approach has been successfully shown to enable the learning agents to observe, learn, and carry out their respective actions as part of the optimal joint action. The LCPP mechanism is a message exchange mechanism that helps the SUs to communicate and compute their own actions.

Our contribution in this paper is to investigate into the LCPP mechanism as an approach to achieve context awareness and intelligence in DCRNs. This approach is potentially useful in addressing existing drawbacks associated with one of the widely-use approaches, namely Game Theory (GT). We show that the LCPP mechanism converges to an optimal joint action including networks with cyclic topology, and fast convergence is possible.

The remainder of this paper is organized as follows. Section II provides a scenario under consideration. Section III discusses research problem and challenges. Section IV presents our system design. Section V presents MARL and our LCPP mechanism. Section VI presents simulation results and discussions. Section VII discusses open issues. Finally, we provide conclusions and future work.

II. A Scenario for DCS in DCRN

In DCS, a problem arises as to what is the best strategy to select an available channel among the licensed channels for data transmission from an SU node given that the objective is to maximize network-wide throughput. The DCRN, as

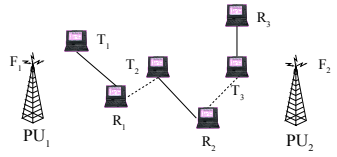

Figure 1. Single-hop DCRN. Solid line indicates communication link; while dotted line indicates interference link.

978-0-7695-4338-3/11 $26.00 © 2011 IEEE
DOI 10.1109/WAINA.2011.47

shown in Figure 1, is modeled using an undirected graph $G(V, E)$, where V is the set of SUs. T_i is the transmitter and R_i is the receiver of an SU communication node pair i; and E is the set of edges between the SUs including the communication links and interference links. There are $U = |V|/2$ SU communication node pairs or agents. We refer to a CR host as a single SU node; and an SU communication node pair as an agent henceforth. Each agent maintains a single set of learned outcomes or knowledge obtained through message exchange; this is necessary because the transmitter and receiver must choose a common channel for data transmission in DCS. An interference edge E_I exists between non-communicating agents if they are located within transmission range of each other. The agents are distributed in a uniform and random manner in a square region. There are K PUs, $PU = [PU_1, \ldots, PU_K]$ and each of them uses one of the K distinctive channels of frequency $F = [F_1, \ldots, F_K]$. In Figure 1, $K = 2 < U = 3$. Each channel frequency is characterized by various levels of PU Utilization Level (PUL) and Packet Error Rate (PER); thus, we consider heterogeneous channels. For a particular channel, higher levels of PUL indicate higher levels of PU activity and hence lesser amount of white spaces; while higher levels of PER indicate higher levels of packet drop rate due to interference, channel selective fading, path loss, and other factors. Spatial reuse is possible where multiple agents are allowed to share a particular channel. The agents infer PUL and PER in each channel in a distributed manner, and select a channel for data transmission individually in order to achieve an optimal joint action so that network-wide throughput is maximized.

III. Research Problem and Challenges

Game Theory has been the most popular approach for achieving context awareness and intelligence in CR networks. GT studies the interaction of multiple SUs whose objective is to maximize their individual local rewards. To date, research has been focusing on one-shot or repetitive games, such as the matrix game, potential game, etc. and these games have been successfully applied in various applications as shown in [4] and [5]. There are several known issues in GT:

- Mis-coordination where the SUs are not able to converge to an optimal joint action because of severe negative rewards [6]. The SUs converge to a *safe* joint action instead.
- The SUs might converge to a sub-optimal joint action when multiple optimal joint actions exist [6].
- GT as applied to CR so far requires a complete set of information to compute the Nash equilibrium in DCRN; hence its extensive and successful usage in centralized CR networks.
- GT assumes that all SUs react rationally as game theorists.
- GT assumes a single type of utility function throughout the DCRN, and hence homogeneous learning mechanism in all the SUs.

Although GT has been successfully applied in CR networks [4], [5], MARL is a good alternative which addresses the aforementioned issues associated with GT [3]. However, it should be noted that MARL is a new research area [7], hence many open issues in MARL are yet to be addressed including the LCPP mechanism.

In a multi-agent environment, there are three main challenges. Firstly, an agent's action is dependent on the other payoff-optimizing agents' actions. Secondly, all agents must converge to an optimal joint action that provides good network-wide performance. Thirdly, all agents must infer the channel heterogeneous characteristics including PUL and PER that might be different among the agents. In other words, from the perspective of each agent, the challenge is "*How does an agent infer its channel characteristics and choose its own action such that the joint action converges to an optimal joint action?*" The traditional single-agent-based reinforcement learning approach enables each agent to learn and choose its action in a unilateral fashion regardless of its neighbour agents' actions in a multi-agent environment, and this may cause instability or oscillation because it switches its action from time to time [3], hence we choose to use MARL. In this paper, we show that, using the LCPP mechanism, which is an important component in MARL, the SUs converge to an optimal joint action in respect to DCS in a distributed manner in DCRN. Our focus in this paper is the payoff message exchange mechanism. In our solution, a local learning mechanism, such as MARL, is available at each SU. The local learning mechanism provides each agent the local reward that an agent could gain for choosing an action; further explanation is provided in the next section.

IV. System Design

In this paper, we model each SU communication node pair as a learning agent, as shown in Figure 2, because the transmitter and receiver share a common learned outcome or knowledge. At a particular time instant, the agent observes its local operating environment in the form of a local reward. Due to the limited sensing capability at each agent, it can only observe its own *local* reward. The agent improves the *global* reward in the next time instant through carrying out a proper action. The MARL is comprised of two components: Local Learning (LL) and the LCPP mechanism. The LL mechanism provides knowledge on the operating environment comprised of multiple agents through observing the consequences of its prior action in the form of local reward [3]. The LCPP mechanism is a message exchange mechanism to help the learning engine embedded

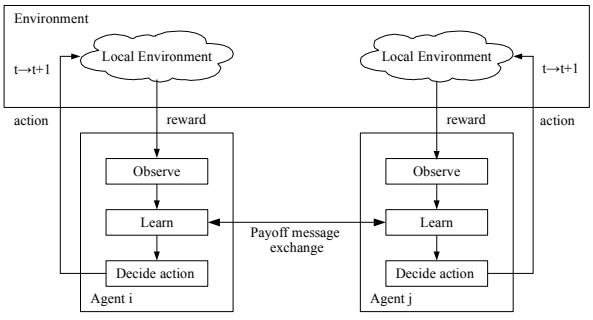

Figure 2. Agents (or SU communication node pairs) and their environment.

in each agent to communicate and compute its own action as part of the optimal joint action using the knowledge provided by the LL mechanism. In other words, the LCPP mechanism is a means of communication for the LL mechanism embedded in each agent. As time progresses, the agents learn to carry out the proper action to maximize the accummulated global rewards. In DCS, the LL mechanism is used to learn the channel heterogeneous characteristics, which is the PUL and PER levels. The global reward is a linear combination or summation of all the local rewards at each agent; while the global payoff is the equivalent for local payoffs generated by each SU. The LCPP mechanism maximizes both the global reward and global payoff in order to achieve an optimal joint action. Based on an application, such as DCS, the reward and payoff values indicate distinctive network performance metrics such as throughput and successful data packet transmission rate. Thus, maximizing the global reward and the global payoff provide network-wide performance enhancement.

V. The MARL Approach

We first describe the Coordination Graph (CG); followed by the LL mechanism, and finally the LCPP mechanism. This paper discusses MARL and PP mechanism in our context and application scenario, and the reader is referred to [3] and [7] for more details.

A. Coordination Graph

In Figure 3, there are $U=|V|/2=4$ agents in the DCRN, and each agent (i.e. T_1 - R_1 pair) is represented by a single node. An interference edge E_I exists between a pair of neighbouring agents. The graph G can be decomposed into smaller and local CGs which are each a local view of the G for each agent. For instance, the CG of agent 1 is comprised of agents 1, 2, and 3, hence the representation of the local reward or Q-value $Q_{i,t}(a_{i,t}, a_{j\in\Gamma(i),t}) = Q_{1,t}(a_{1,t}, a_{2,t}, a_{3,t})$, where $a_i \in A$. The CG defines collaborative relationships, and each relationship is a local payoff message exchange mechanism among the agents. Each agent runs an LL mechanism independently to update its own Q-values. The approximate global Q-value $Q_t(\boldsymbol{a}_t)$ at time t is a linear combination or summation of all the local Q-values at each agent as follows:

$$Q_t(\boldsymbol{a}_t)=\sum_{i=1}^{U} Q_{i,t}(a_{i,t}, a_{j\in\Gamma(i),t}) \quad (1)$$

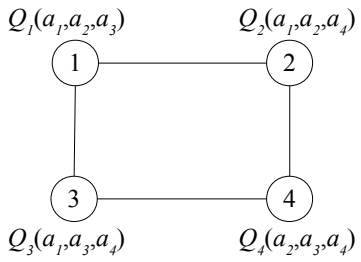

Figure 3. A four-agent graph G. Each agent represents an SU communication node pair. The edges are interference edges that exist between the agents that interfere with each other.

where $\Gamma(i)$ represents all the neighbours of agent i.

B. Local Learning Mechanism

The Q-value $Q_{i,t}(a_{i,t}, a_{j\in\Gamma(i),t})$, which is the learned knowledge and maintained in a lookup Q-table with $|A|$ entries at agent i, represents the local reward that the agent can gain for choosing an action $a_i \in A=F$. For example, in DCS, the Q-value represents the throughput and it is dependent on the local PUL, PER and the joint action $\boldsymbol{a}_t$ taken by all the agents. The joint action affects the Q-value due to the dependency of actions among the agents; for example, two neighbour agents that choose a particular action, specifically a channel, might increase their contention level, and hence reduce their respective Q-values for the action.

C. Locally-Confined Payoff Propagation (LCPP) Algorithm

The pseudo-code of our LCPP algorithm is shown in Algorithm 1, and it is embedded in each agent. Each agent i constantly sends reward value or payoff message $\mu_{ij}(a_{j,t})$ to its neighbour agents $j\in\Gamma(i)$ over the edges as shown in Figure 4. Channel selection is based on an agent's two-hop neighbour agents in DCRNs, hence each $\mu_{ij}(a_{j,t})$ contains two pieces of information: 1) its own Q-value; and 2) its one-hop neighbours' local Q-value except that from agent j, as follows:

$$\mu_{ij}(a_{j,t})=[Q_{i,t}(a_{i,t}, a_{j\in\Gamma(i),t}), \sum_{k\in\Gamma(i)\setminus j} Q_{k,t}(a_{k,t}, a_{l\in\Gamma(k),t})] \quad (2)$$

where $\Gamma(i)\setminus j$ represents all the neighbours of agent i except agent j. Using $\mu_{ij}(a_{j,t})$, agent i informs agent j about the Q-values of itself and its neighbour agents when agent j is taking its own action so that agent j can evaluate its own action.

The payoff messages are exchanged among the agents until convergence to a fixed optimal point occurs within a finite number of iterations. Before convergence, the messages are an estimation of the fixed optimal point as all incoming messages are yet to converge. Each agent selects its own optimal action, which is part of the optimal joint action, to maximize the local payoff as shown next:

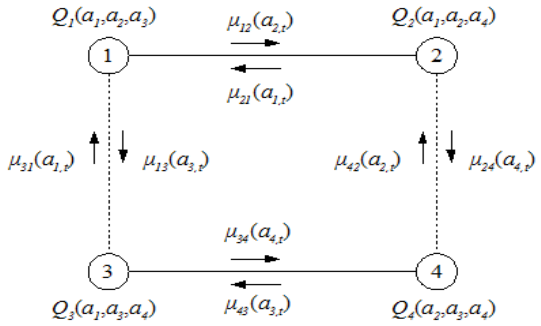

Figure 4. Illustration of four agents and payoff message exchange.

$$g_i(a_{i,t}) = \max_{a \in A} [Q_{i,t}(a, a_{j \in \Gamma(i),t}) + \sum_{j \in \Gamma(i)} Q^i_{j,a,t}(a_{j,t}, a_{k \in \Gamma(j),t}) + \sum_{k \in \Gamma(j) \setminus j} Q^i_{k,a,t}(a_{k,t}, a_{l \in \Gamma(k),t})] \quad (3)$$

where $Q^i_{j,a,t}$, which is kept at agent i, indicates the Q-value of agent j when agent i is taking action a.

Each agent i determines its optimal action individually as follows:

$$a^*_{i,t} = \underset{a \in A}{argmax}\, g_i(a) \quad (4)$$

The approximate global payoff $g(\boldsymbol{a_t})$ at time t is a linear combination or summation of all local payoff as follows:

$$g(a_t) = \sum_{i=1}^{U} [Q_{i,t}(a_i, a_{j \in \Gamma(i),t}) + \sum_{j \in \Gamma(i)} Q^i_{j,a_i,t}(a_{j,t}, a_{k \in \Gamma(j),t}) + \sum_{k \in \Gamma(j) \setminus j} Q^i_{k,a_i,t}(a_{k,t}, a_{l \in \Gamma(k),t})] \quad (5)$$

The agents would reach a fixed optimal point after a finite number of iterations. Note the difference between the global Q-value, $\Sigma_i Q_i$ in (1) and the global payoff, $\Sigma_i [Q_i + \Sigma_j \mu_{ji}]$ in (5). The global Q-value is the total rewards received by all the agents in the network; while the global payoff is the total local Q-value and local payoff value exchanged among the agents. Both global Q-value and global payoff are the performance metrics for the LCPP mechanism. They converge to an optimal joint action. Both equations (1) and (5) are the performance metrics for the LCPP algorithm.

```
initialize μ_ij = μ_ji = 0 for j ∈ Γ(i) , g_i = 0
{Task: 1. Broadcast payoff message to neighbour agents
       2. Select optimal action}
if ( my turn to select an optimal action )
   R = uniform(0,1)
   if ( R ≤ ε ) then
      a = uniform(1,K) {Select random action}
   else
      for ( all neighbour agents )
         compute μ_ij(a_j,t) {Equation (2)}
         if ( μ_ij(a_j,t) ≠ μ_ij(a_j,t-1) ) then
            include μ_ij(a_j,t) in broadcast message
            compute a = a*_i,t  {Select optimal action}
                                {Equation (3) and (4)}
         end if
      end for
   end if
   return a
end if
{Task: 3. Process payoff message}
if ( receive message )
   if ( message = μ_ij(a_j,t) ) then store this information
   end if
end if
```

Algorithm 1. Pseudo-code of the LCPP algorithm at agent *i*.

Selecting the optimal action for all the times does not cater for the actions that are never chosen. Exploitation chooses the optimal action. Exploration chooses the other possible actions in A in order to improve the estimates of the other Q-values. In the ε-greedy approach [8], an agent performs exploration with small probability ε, and exploitation with probability $1-\varepsilon$.

D. Difference between the Original PP mechanism and the Locally-Confined PP mechanism

The original PP mechanism [3] has been shown to be useful in many applications, however, our LCPP mechanism provides two advantages so that it is suitable to be applied in DCS in DCRN. Firstly, LCPP is suitable for DCRNs with cyclic topology because old payoff messages are not added to the current payoff message. As explained in [3], the payoff message $\mu_{ij}(a_{j,t})$ increases without bound in a cyclic topology if old payoff values are added into current payoff value computation. The original PP mechanism requires the agents to compute global payoff in the entire network from time to time to alleviate this issue. Since old payoff messages are not added to the current payoff message in LCPP, our simulation results show that the global payoff value does not increase without bound. Hence, LCPP does not require the agents to compute global payoff in the entire network from time to time. Secondly, in DCS, an agent selects its channel based on the channel selections of its two-hop neighbour agents. An agent uses Q-values from its two-hop neighbour agents only in payoff value computation in LCPP, while in [3], an agent uses Q-values from agents that are multiple hops away.

The LCPP mechanism provides modification to the original PP mechanism so that it is suitable to be applied in DCRN. Since the original PP mechanism cannot be directly applied in DCRN, we do not investigate into performance comparison between the original PP mechanism and the LLCP mechanism.

VI. Simulation Results and Discussions

Our objective is to enable the SUs to select their channel (action) for data transmission respectively in DCS. The channel selections (joint action) by all the SUs provide an optimal network-wide throughput (global Q-value or global reward). In this paper, simulations were performed using C programming, rahter than network simulator, so that the simulation results are not dependent on the underlying Medium Acccess Control (MAC) protocol.

The simulation scenario is discussed in Section II. Graphical representation of the DCS scheme is shown in Figure 5. We perform simulation using $G=(V,E)$ with $U=|V|/2=10$ and three levels of densities with different number of interference links in the entire network $|E_I|=\{Low=5, Medium=10, High=15\}$, and cyclic

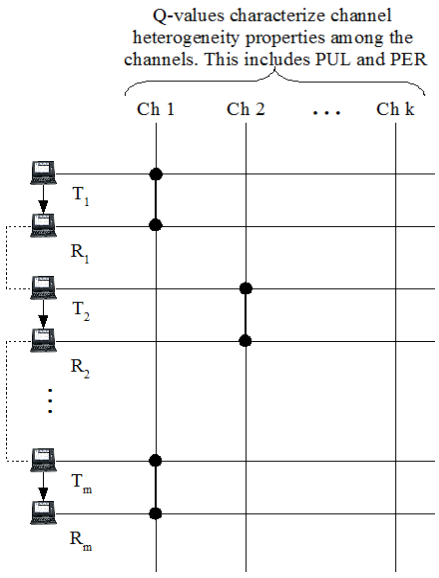

Figure 5. Graphical representation of the DCS scheme. Bold line indicates data transmission between a communication node pair (an agent) over a chosen channel; while dotted line indicates interference link.

topology exists. There are $K=3$ channels. The Q-value characterizes the channel heterogeneity properties for each channel including PUL and PER. For a particular channel, the Q-values are different among the agents as each of them observes different levels of PUL and PER. In short, the Q-values are Independent and Identically Distributed (i.i.d.) among the agents and the channels. In the simulation, each Q-value has a range $-5 \leqslant Q_i(a_i, a_{j\in\Gamma(i)}) \leqslant 15$. Higher Q-value indicates better reward, and hence higher throughput. At each iteration t, communication node pair i chooses a channel out of $K=3$ channels for data transmission, and it explores with probability $\varepsilon=0.0125$. A single iteration corresponds to the conventional four-way handshaking mechanism that covers payoff message exchange in RTS and CTS, data packet and ACK transmission.

Figure 6 shows that the global payoff of the low, medium and high density networks increases and converges to a

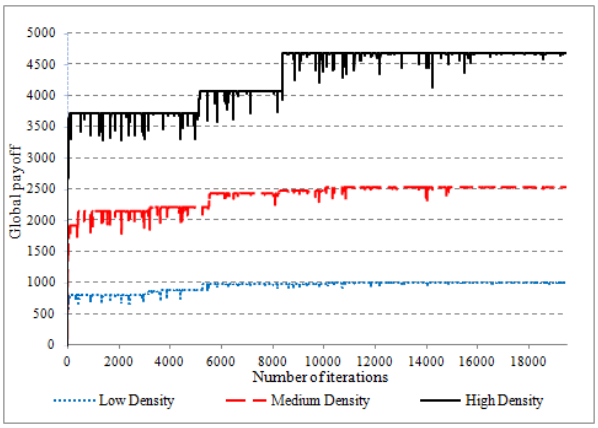

Figure 6. Global payoff for low, medium and high density networks.

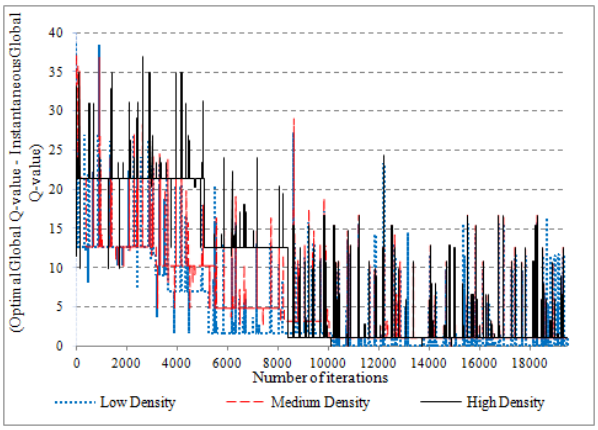

Figure 7. Difference between optimal global Q-value and instantaneous global Q-value for low, medium and high density networks.

fixed point as the number of iterations advances. The high-density network converges in approximately 9000 iterations, while medium and low-density networks converge in approximately 5000 iterations. The global payoff fluctuates once in a while due to occasional exploration. The global payoff converges in cyclic topology. The high-density network has the highest global payoff because the payoff depends on the number of neighbour agents.

Figure 7 shows that the difference between the optimal and instantaneous global Q-values of the low, medium and high density networks decreases to approximately zero. This indicates the convergence to an optimal joint action. The instantaneous global Q-value is calculated using (1); and the optimal global Q-value is the linear combination or summation of all optimal local Q-values at each agent. The difference value fluctuates once in a while due to occasional exploration.

Figure 8 shows that the global payoff of a medium-density network increases and converges to a fixed point as the number of iterations advances in the presence of varying Q-values for the first 2000, 4000 and 6000 iterations. At each iteration, each Q-value varies with probability 0.5. The

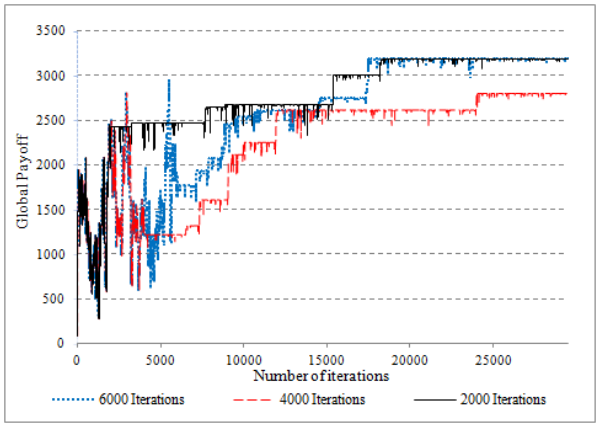

Figure 8. Global payoff for medium density network with varying Q-values for the first 2000, 4000 and 6000 iterations.

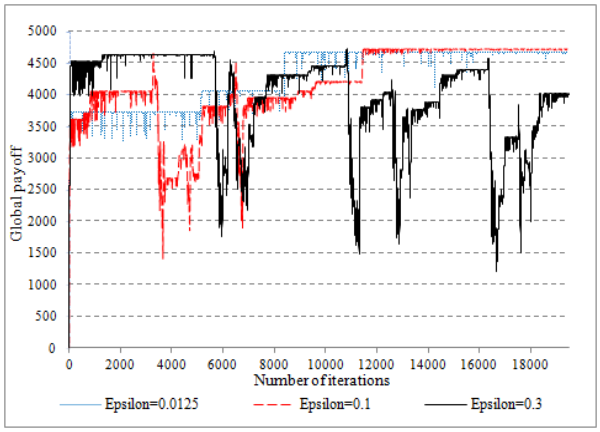

Figure 9. Global payoff for different ε in a high-density network.

results indicate that stable Q-values provided by the LL mechanism is the key factor for convergence.

Figure 9 shows that the global payoff of the high density network increases and converges to a fixed point as the number of iterations advances. The network converges in approximately 9000 iterations when $\varepsilon=0.0125$, 11000 iterations when $\varepsilon=0.1$, and 50 iterations when $\varepsilon=0.3$. Thus, higher ε values improve convergence rate. However, the global payoff fluctuates when $\varepsilon=0.1$ and $\varepsilon=0.3$ due to excessive exploration. This is why $\varepsilon=0.0125$ has been chosen in our simulations although it has the lowest convergence rate. This key finding shows that it is possible to achieve fast convergence using high values of ε and subsequently to achieve stability using lower values of ε upon reaching the stability state.

VII. Open Issues

The open issues are relevant to the MARL, LL and LCPP mechanisms respectively. MARL is a new reesarch area [7] suitable for tackling the drawbacks of GT. These open issues are subject to further investigation. In MARL, it is necessary to detect and respond to irrational agents that take random or suboptimal actions. In LL, there are two issues. Firstly, mechanism to provide stable Q-values. This includes detecting and responding to any fluctuations in the Q-values. This is critical so that the LCPP mechanism converges to an optimal joint action (see Figure 8). Secondly, mechanism for heterogeneous learning among the agents. Each agent represents the Q-values with different performance metrics in a particular DCRN to enable heterogeneous learning entities. In LCPP, there are four issues. Firstly, mechanism to improve the convergence rate of the global payoff (see Figure 6) and global Q-value (see Figure 7). Secondly, investigation into the impact of dynamics in the operating environment on ε, as well as mechanism to measure the level of dynamics for dynamic adjustment of ε. Thirdly, investigation into tradeoff between convergence rate and stability through the adjustment of ε. Fourthly, comparison of the amount of overhead between the LCPP and GT approaches.

Conclusions and Future Work

In this paper, we have presented our Locally-Confined Payoff Propagation (LCPP) mechanism, which is an important feature in the Multi-Agent Reinforcement Learning (MARL) approach to achieve optimal joint action in a Distributed Cognitive Radio Network (DCRN). We have shown that the LCPP mechanism converges to an optimal joint action including networks with cyclic topology. Fast convergence is possible through the adjustment of the exploration probability ε. In our future work, we will investigate the open issues raised in this paper. The investigations in this paper serve as the basis for future research in DCRN.

References

[1] FCC Spec Plcy Tsk Frc, "Report of the Spectrum Efficiency Working Group," *Fed Comm Comsn, Tech Rpt 02-155*, US, Nov. 2002.

[2] Mitalo III, J., and Maguire, G. Q., "Cognitive radio: Making software radios more personal," *IEEE Psnl. Comm.*, 6, pp. 13-18, 1999.

[3] J. R. Kok, and N. Vlassis "Collaborative Multiagent Reinforcement Learning," *J. Mach. Learn. Research 7*, pp. 1789-1828, Sep. 2006.

[4] I. Malanchini, M. Cesana, and N. Gatti, "On Spectrum Selection Games in Cognitive Radio Networks," *IEEE Global Telecom. Conf. (GLOBECOM)*, Honolulu, HI, Dec 2009.

[5] H. Li, and Z. Han, "Competitive Spectrum Access in Cognitive Radio Networks: Graphical Game and Learning," *IEEE Wls. Comm. and Nwk. Conf. (WCNC)*, Sydney, Australia, April 2010.

[6] S. Kapetanakis, D. Kudenko, and M. J. A. Strens, "Reinforcement learning approaches to coordination in cooperative multi-agent systems," *Adptv. Ag. and Multi-ag. Sys,* LNCS, Springer Berlin, 2003.

[7] L. Busoniu, R. Babuska, and B. D. Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Trans. on Sys, Man, & Cys, Part C: Appl. & Rew.*, 38(2), pp. 156-172, Mar. 2008.

[8] R. S. Sutton and A. G. Barto, Reinforcement learning: an introduction. Cambridge MA, MIT Press, 1998.