
Article:
Morozs, Nils, Clarke, Tim and Grace, David (2016) Distributed Heuristically Accelerated Q-Learning for Robust Cognitive Spectrum Management in LTE Cellular Systems. IEEE Transactions on Mobile Computing. pp. 817-825.

https://doi.org/10.1109/TMC.2015.2442529

Distributed Heuristically Accelerated Q-Learning for Robust Cognitive Spectrum Management in LTE Cellular Systems

Nils Morozs, *Student Member, IEEE,* Tim Clarke, and David Grace, *Senior Member, IEEE*

Abstract—In this paper we propose an algorithm for dynamic spectrum access (DSA) in LTE cellular systems - distributed ICIC accelerated Q-learning (DIAQ). It combines distributed reinforcement learning (RL) and standardized inter-cell interference coordination (ICIC) signalling in the LTE downlink, using the framework of heuristically accelerated RL (HARL). Furthermore, we present a novel Bayesian network based approach to theoretical analysis of RL based DSA. It explains a predicted improvement in the convergence behaviour achieved by DIAQ, compared to classical RL. The scheme is also assessed using large scale simulations of a stadium temporary event network. Compared to a typical heuristic ICIC approach, DIAQ provides significantly better quality of service and supports considerably higher network throughput densities. In addition, DIAQ dramatically improves initial performance, speeds up convergence and improves steady state performance of a state-of-the-art distributed Q-learning algorithm, confirming the theoretical predictions. Finally, our scheme is designed to comply with the current LTE standards. Therefore, it enables easy implementation of robust distributed machine intelligence for full self-organisation in existing commercial networks.

Keywords—Heuristically Accelerated Q-Learning, Dynamic Spectrum Access, Inter-Cell Interference Coordination.

1 Introduction

One of the fundamental tasks of a cellular system is spectrum management, concerned with dividing the available spectrum into a set of resource blocks and assigning them to voice calls and data transmissions in a way which would provide a good quality of service (QoS) to the users. Flexible dynamic spectrum access (DSA) techniques play a key role in utilising the given spectrum efficiently. For example, one of the key requirements for LTE systems is to have a reuse factor of 1 [1]. Therefore, there is an inherent need for DSA techniques in such systems to mitigate the effects of inter-cell interference on the system throughput and the QoS provided to the mobile subscribers. In order to achieve this, LTE systems use a dedicated X2 interface for exchanging relevant interference information among neighbouring eNodeBs (eNBs) [2]. This process is referred to as inter-cell interference coordination (ICIC).

An emerging state-of-the-art technique for intelligent DSA is reinforcement learning (RL); a machine learning technique aimed at building up solutions to decision problems only through trial-and-error [3]. It has been successfully applied to a range of DSA problems and scenarios, such as cognitive radio networks [4], femtocell networks [5], cognitive wireless mesh networks [6], as well as generic cellular networks [7]. The most widely used RL algorithm in both artificial intelligence and wireless communications domains is Q-learning [8]. Therefore, most of the literature on RL based DSA focuses on Q-learning and its variations, e.g. [6][7][9]. This paper investigates distributed Q-learning based DSA. The distributed Q-learning approach has advantages over centralised methods in that no communication overhead is required to achieve the learning objective, and the network operation does not rely on a single computing unit. It also allows for easier insertion and removal of base stations from the network, if necessary. For example, such distributed opportunistic protocols are well suited to temporary event networks and disaster relief scenarios, where rapidly deployable network architectures with unplanned or variable topologies may be required to supplement any existing wireless infrastructure [10].

Although RL algorithms such as Q-learning have been shown to be a powerful approach to problem solving, their common disadvantage is the need for many learning iterations to converge on an acceptable solution. One of the more recent promising solutions to this issue, proposed in the artificial intelligence domain, is the heuristically accelerated reinforcement learning (HARL) approach. Its goal is to speed up RL algorithms, particularly in the multi-agent domain, by guiding the exploration process using additional heuristic information [11]. In [12], case-based reasoning is used for heuristic acceleration in a multi-agent RL algorithm to assess similarity between states of the environment and to make a guess at what action needs to be taken in a given state, based on the experience obtained in other similar states. In [11], Bianchi et al. prove the convergence of four multi-agent HARL algorithms and show how they outperform the regular RL algorithms. There is no evidence in the literature of the HARL approach being applied in the wireless communications domain.

The purpose of this paper is to resolve the problem of

poor temporal performance of RL based DSA algorithms, by proposing a cognitive DSA scheme which combines distributed Q-learning and ICIC using a novel adaptation of the HARL framework. Furthermore, it is designed to comply with the current LTE standards and enables robust distributed machine intelligence to be easily implemented in current or future LTE releases.

In previous work on combining ICIC and RL, researchers have only considered applying RL to learning various parameters related to ICIC or radio resource management in OFDMA cellular systems, such as LTE or WiMAX. For example, Simsek et al. [13] use RL to learn optimal cell range bias and power allocation strategies and compare them to static ICIC methods; Dirani and Altman [14] use a fuzzy Q-learning algorithm and ICIC to learn a coordinated power allocation strategy; and Vlacheas et al. [15] use a fuzzy RL principle for automatic tuning of the Relative Narrowband Transmit Power (RNTP) indicator, which is a key ICIC parameter in the LTE downlink. However, there is no evidence of previous work in the literature on using heuristic ICIC methods to enhance the performance of RL based DSA algorithms.

The rest of the paper is organised as follows: Section 2 explains the current specification of ICIC signalling in the LTE downlink. Section 3 describes the distributed Q-learning approach to DSA. In Section 4 we introduce our formulation of the HARL framework and propose the distributed ICIC accelerated Q-learning (DIAQ) scheme for DSA in LTE cellular systems. In Section 5 we use a novel Bayesian network based method for theoretical analysis and evaluation of the proposed scheme. Section 6 evaluates its performance by simulating a large scale cellular system. The conclusions are given in Section 7.

2 Inter-Cell Interference Coordination in LTE Downlink

The main limiting factor for network throughput performance in LTE systems is inter-cell interference, since the key requirement for LTE systems is to have a reuse factor of 1 [1], i.e. the full spectrum pool is available to every eNodeB (eNB) in the network. The same applies to other future cellular systems which will employ advanced DSA techniques, as opposed to static resource allocation methods [16]. Consequently, the key interference management technology investigated within the context of LTE is inter-cell interference coordination (ICIC). The purpose of ICIC is to reduce interference between adjacent cells by exchanging information between neighbouring eNBs over the X2 interface [1]. This ICIC signal exchange is depicted in Fig. 1 using a generic hexagonal cell network architecture. Here, the central eNB is sending an ICIC signal to the eNBs around it to let them know in which parts of the spectrum it is likely to interfere with them.

The format of the messages exchanged between eNBs using ICIC in the LTE downlink is standardized by

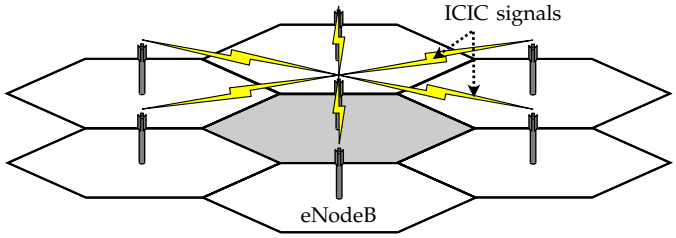


Fig. 1. ICIC signalling among adjacent eNodeBs

the 3GPP and referred to as the Relative Narrowband Transmit Power (RNTP) indicator [17]. It contains a bitmap which indicates on which resource blocks an eNB is planning to transmit at high power by setting their corresponding bits to 1. The threshold used to decide if a transmit power is high or low is derived using the RNTP threshold, which can take the following set of standardized values:

$$RNTP_{threshold} \in \{\infty, -11, -10, ..., 3\}\ dB \quad (1)$$

It is measured in dB relative to the average transmit power in a given cell.

3 Distributed Q-Learning Based Dynamic Spectrum Access

In pure distributed reinforcement learning (RL) based DSA the task of every eNB is to learn to prioritise among the available subchannels only through trial-and-error, with no frequency planning involved, and with no information exchange with other eNBs, e.g. [7]. In this way, frequency reuse patterns emerge autonomously using distributed artificial intelligence with no requirement for any prior knowledge of a given environment.

3.1 Reinforcement Learning

RL is a model-free type of machine learning which is aimed at learning the desirability of taking any available action in any state of the environment only through trial-and error [3]. This desirability of an action is represented by a numerical value known as the Q-value - the expected cumulative reward for taking a particular action in a particular state, as shown in the equation below:

$$Q(s,a) = E\left[\sum_{t=0}^{T} \gamma^t r_t\right] \quad (2)$$

where $Q(s,a)$ is the Q-value of action a in state s, r_t is the numerical reward received t time steps after action a is taken in state s, T is the total number of time steps until the end of the learning process or episode, and $\gamma \in [0,1]$ is a discount factor.

The job of an RL algorithm is to estimate $Q(s,a)$ for every action in every state, which are then stored in an array known as the Q-table. In some cases where an

environment does not have to be represented by states, only the action space and a 1-dimensional Q-table $Q(a)$ can be considered [18]. The job of an RL algorithm then becomes simpler; it aims to estimate an expected value of a single reward for each action available to the learning agent:

$$Q(a) = E[r_t] \tag{3}$$

This is also applicable to distributed Q-learning based DSA in cellular systems, e.g. [7][19][20].

3.2 Distributed Stateless Q-Learning

One of the most widely used RL algorithms is Q-learning [8]. In particular, a simple stateless variant of this algorithm, as formulated in [18], has been shown to be effective for several distributed DSA learning problems, e.g. [7][19][20].

Each eNB maintains a Q-table $Q(a)$ such that every subchannel a has a Q-value associated with it. Upon each file arrival, the eNB either assigns a subchannel to its transmission or blocks it if all subchannels are occupied. It decides which subchannel to assign based on the current Q-table and the greedy action selection strategy described by the following equation:

$$\hat{a} = \underset{a}{\operatorname{argmax}}(Q(a)) \tag{4}$$

where $\hat{a}$ is the subchannel chosen for assignment, and $Q(a)$ is the Q-value of subchannel a.

The values in the Q-tables are initialised to zero, so all eNBs start learning with equal choice among all available subchannels. A Q-table is updated by an eNB each time it attempts to assign a subchannel to a file transmission in the form of a positive or a negative reinforcement. The recursive update equation for stateless Q-learning, as defined in [18], is given below:

$$Q(a) \leftarrow (1-\alpha)Q(a) + \alpha r \tag{5}$$

where $Q(a)$ represents the Q-value of the subchannel a, r is the reward associated with the most recent trial and is determined by a reward function, and $\alpha \in [0, 1]$ is the learning rate parameter which weights recent experience with respect to previous estimates of the Q-values.

The reward function, which is generally applicable to a wide range of RL problems and which has been successfully applied to DSA problems in the past [4][7], returns two values:

- $r = -1$ (negative reinforcement), if the file transmission failed due to an insufficient SINR on the selected subchannel.
- $r = 1$ (positive reinforcement), if the file is successfully transmitted, i.e. SINR did not drop below the transmission threshold.

The choice of the learning rate values for this type of distributed Q-learning based DSA problems is thoroughly investigated in [7]. The best performance is achieved by using the Win-or-Learn-Fast (WoLF) principle [21] described by (6), where a lower value of α is used for successful trials (when $r = 1$), and a higher value of α is used for failed trials ($r = -1$). In this way, the learning agents learn faster when "losing" and more slowly when "winning".

$$\alpha = \begin{cases} 0.01 & r = 1 \\ 0.05 & r = -1 \end{cases} \tag{6}$$

4 Distributed Heuristically Accelerated Q-Learning

A common disadvantage of machine learning algorithms, such as distributed Q-learning described in the previous section, is that they are normally used to learn solutions only through trial-and-error with no prior knowledge of the problem in hand. Consequently, it takes a large number of trials for them to learn acceptable solutions. This is undesirable in real-time applications such as DSA in cellular systems. An emerging technique to mitigate this poor initial performance problem is the heuristically accelerated reinforcement learning (HARL) approach, where additional heuristic information is used to guide the exploration process [11].

4.1 Heuristically Accelerated Reinforcement Learning

The key additional element provided by HARL compared to classical RL is the derivation of a heuristic function. According to [11], a heuristic function is derived from additional knowledge, either external or internal, which is not included in the learning process. Generally, the goal of the heuristic function $H_t(s, a)$ is to influence the action choices of a learning agent, i.e. to modify its current policy $\pi_t(s)$ in a way which would accelerate the learning process. The format and dimensions of $H_t(s, a)$ should be compliant with the Q-table used by the given learning agent, such that its new combined policy $\pi^c_t(s)$ can be derived using the following equation:

$$\pi^c_t(s) = \underset{a}{\operatorname{argmax}}(Q_t(s, a) + H_t(s, a)) \tag{7}$$

where $\pi^c_t(s)$ is the combined policy of the given learning agent for state s at time t based on its Q-table $Q_t(s, a)$ and the heuristic function $H_t(s, a)$. If $H_t(s, a)$ is always zero, the algorithm becomes a regular Q-learning algorithm with a greedy action selection strategy. In the case of the stateless Q-learning algorithm described in Section 3, the heuristic function would not have a state dimension and can be denoted by $H_t(a)$.

4.2 Distributed ICIC Accelerated Q-Learning

In this subsection we propose the distributed ICIC accelerated Q-learning (DIAQ) DSA scheme that combines distributed Q-learning and ICIC using the HARL framework to mitigate the issue of poor temporal performance characteristics of Q-learning based DSA algorithms.

As described in Section 2, by using ICIC signalling over the X2 interface, every eNB has the capability of knowing on which virtual resource blocks (VRBs) the neighbouring eNBs are likely to interfere with it, i.e. transmit at a power above an RNTP threshold. In a scenario, where a 20 MHz LTE channel consisting of 100 VRBs is allocated to the network, the length of an RNTP message is 100 bits or 25 hexadecimal characters. There, every subchannel, i.e. a minimum entity allocated to a file transmission, consists of 4 adjacent VRBs, if resource allocation "Type 0" is used [17]. In case of the central eNB in Fig. 1, it receives 6 RNTP messages from its neighbours, each containing 25 hexadecimal characters, stating subchannels they need to reserve to avoid inter-cell interference. $0xF$ denotes that a subchannel is in use by the neighbouring eNB, and $0x0$ means it is safe to use by the eNB which receives the RNTP message.

We propose using these RNTP messages for creating ICIC bitmasks indicating which subchannels are not safe to use for any given eNB, as notified by its neighbours, and using these bitmasks for creating heuristic functions $H(a)$, which in turn influence the spectrum assignment choices made by the distributed Q-learning based DSA algorithm.

When a request for a new file transmission is received, the eNB starts by aggregating the latest RNTP messages from its neighbours into an ICIC bitmask using a bitwise *OR* operation, as described by the following equation:

$$Mask = \bigcup_{n=1}^{N} RNTP_n \tag{8}$$

where $Mask$ is a 25 hexadecimal character string representing the subchannels reserved by any of the neighbouring base stations by F, and representing the "safe-to-use" subchannels by 0, $RNTP_n$ is a 25 hexadecimal character RNTP message of the n'th neighbouring eNB, and N is the total number of neighbouring eNBs. The RNTP message exchanges can take place as often as every 20 ms [1], and they do not have to be synchronised. Every eNB always uses the latest RNTP message received from a given neighbour.

After creating the ICIC mask, we the eNB derives a heuristic function $H(a)$ as follows:

$$H(a) = \begin{cases} h & Mask(a) = F \\ 0 & Mask(a) = 0 \end{cases} \tag{9}$$

where $H(a)$ is the value of the heuristic function for subchannel a, and h is a fixed negative value with a greater amplitude than the full range of possible $Q(a)$ values. In case of the distributed Q-learning algorithm described in Section 3, $Q(a) \in [-1, 1]$, therefore $h < -2$. $H(a)$ can be employed to create a temporary masked Q-table $Q_m(a)$ using the following equation:

$$Q_m(a) = Q(a) + H(a) \tag{10}$$

$Q_m(a)$ is then used for heuristically guided decision making, whilst a normal learning process takes place using the $Q(a)$, as defined in (5).

By using the proposed $Q_m(a)$ and $H(a)$, the eNB is guaranteed to prioritise the subchannels marked as "safe" by $Mask$ before the "unsafe" subchannels by shifting the Q-values of the latter to the bottom of the Q-table, whilst still preserving their respective order in terms of the Q-values (due to the fixed value of h).

The detailed flowchart of the proposed DIAQ scheme is shown in Fig. 2. The novel ICIC related algorithm steps are shaded and use dotted outlines. The rest of the flowchart describes a regular distributed Q-learning based DSA process described in Section 3. The shaded blocks with solid outlines indicate the functions which drive the RL process, i.e. update the Q-table.

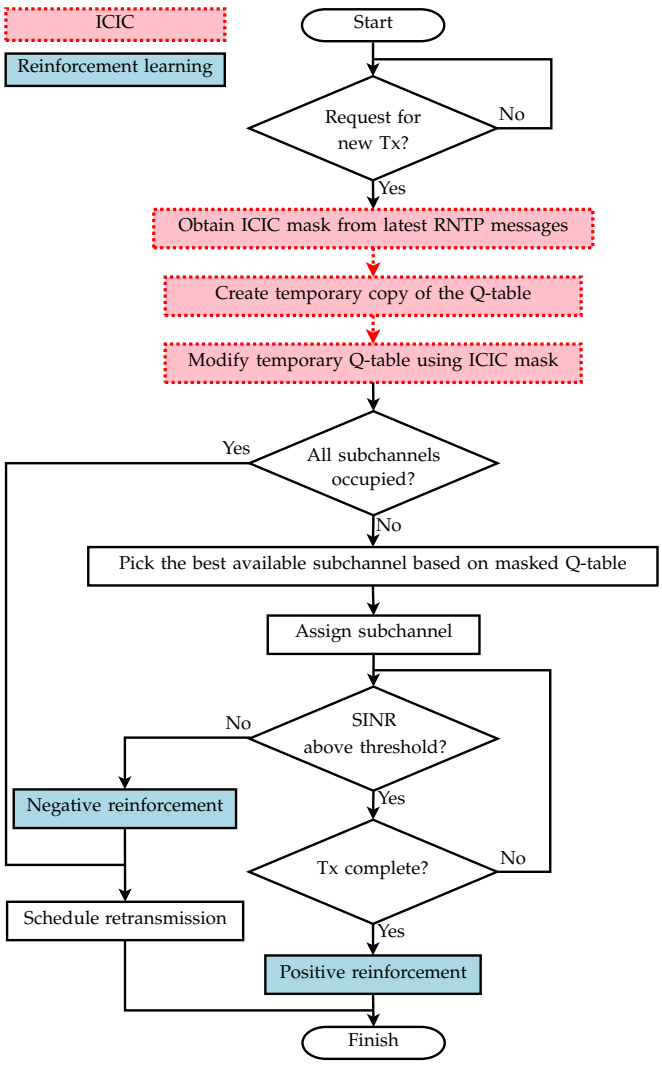


Fig. 2. Flowchart of the proposed distributed ICIC accelerated Q-learning (DIAQ) scheme

5 Theoretical Evaluation

Before testing the developed DIAQ scheme in a realistic scenario, its expected performance improvements over regular distributed Q-learning are analytically derived using a simple model which represents a generalized inter-cell interference problem.

5.1 Simple Inter-Cell Interference Model

The network model used for analysis in this section is depicted in Fig. 3. It consists of 2 eNBs and 2 user equipments (UEs), each connected to its own eNB. If one of the UEs is located within the interference range of the other eNB, it suffers from harmful co-channel interference from it. The network is assumed to be allocated 2 subchannels, and the task of both eNBs is to learn to use their own subchannel through distributed Q-learning and DIAQ.

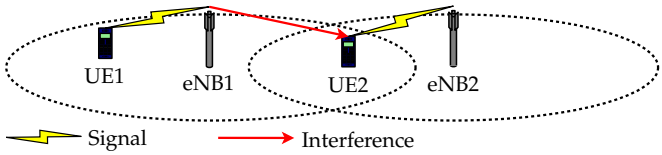


Fig. 3. 2 eNB 2 UE network model

5.2 Bayesian Network Model

Bayesian networks are a powerful tool for modelling conditional dependencies among stochastic variables [22]. Fig. 4 presents a novel Bayesian network model which describes the behaviour of DIAQ when applied to the simple DSA network model from Fig. 3. The shaded nodes and dotted edges show extra dependencies introduced by DIAQ, compared to classical Q-learning from Section 3. The variables used to denote the Bayesian network nodes are the following:

- $RNTP \in \{Yes, No\}$ - whether or not, at the latest file arrival time, the corresponding eNB has an up-to-date RNTP message from its neighbour.

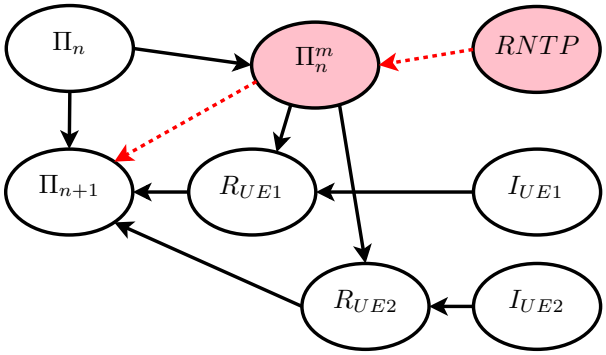


Fig. 4. Bayesian network describing the behaviour of Q-learning and DIAQ

- $I_{UEx} \in \{Yes, No\}$ - whether or not $UE1$ or $UE2$ is located within the interference range of the adjacent eNB during the current file arrival.
- $\Pi_n \in \{Same, Diff\}$ - joint policy of the eNBs after n learning iterations. The policy of an eNB is defined as its preferred subchannel (1 or 2), based on (4). Π_n takes two values of interest - whether the policies of 2 eNBs are the same or different ($\Pi_n = Diff$ is the learning objective).
- $\Pi_n^m \in \{Same, Diff\}$ - joint masked policy, i.e. the combination of Π_n and the heuristic functions of both eNBs defined in (9). It is conditionally dependent on Π_n and $RNTP$ (Π_n^m may be different to Π_n, based on the transformation defined in (10)).
- $R_{UEx} \in \{S, F\}$ - whether or not a file transmission to $UE1$ or $UE2$ was successful (S), or whether it failed (F) due to interference. It is conditionally dependent on Π_n^m and I_{UEx}.
- $\Pi_{n+1} \in \{Same, Diff\}$ - the updated joint policy for the next iteration as a result of the outcome at the current iteration. It is conditionally dependent on Π_n, Π_n^m, R_{UE1} and R_{UE2}.

Based on the conditional dependencies described above and depicted in the Bayesian network model, the equation for calculating the joint probability distribution over all variables $P_{joint} = P(\Pi_{n+1}, \Pi_n, \Pi_n^m, R_{UE1}, R_{UE2}, I_{UE1}, I_{UE2}, RNTP)$ is the following:

$$\begin{aligned} P_{joint} &= P(\Pi_{n+1}|\Pi_n, \Pi_n^m, R_{UE1}, R_{UE2}) \\ &\times P(R_{UE1}|\Pi_n^m, I_{UE1})\ P(R_{UE2}|\Pi_n^m, I_{UE2}) \\ &\times P(\Pi_n^m|\Pi_n, RNTP)\ P(\Pi_n)\ P(RNTP) \\ &\times P(I_{UE1})\ P(I_{UE2}) \end{aligned} \tag{11}$$

which consists of a number of prior probabilities of the form $P(X)$, and conditional probabilities of the form $P(X|Y_1...Y_n)$.

The prior probability distributions that appropriately describe the scenario depicted in Fig. 3 are defined in Table 1. Before any file arrivals at either eNB, the Q-tables of both eNBs are initialised to zero for both subchannels. Therefore, there is a 50% chance of the eNBs choosing the same subchannel, since both of them choose a subchannel at random, i.e. $P(\Pi_0 = Same) = 0.5$. Furthermore, it is assumed without the loss of generality that the interference range overlap of the eNBs is such that there is a 40% chance of a UE being located in it, i.e. $P(I_{UEx} = Yes) = 0.4$. Finally, $P(RNTP = Yes) = High$ represents a high chance of an RNTP message exchange taking place between current file arrivals at the two eNBs. Since these exchanges can take place as often as every 20 ms, an eNB is highly likely to have an up-to-date RNTP message from its neighbour. If $P(RNTP = Yes)$ is changed to 0, the Bayesian network model will describe the Q-learning algorithm from Section 3.

The conditional probability distributions are defined in Table 2. The values used for $P(\Pi_n^m|\Pi_n, RNTP)$ state that the masked policies Π_n^m of the eNBs will be the same ($Same$) with a probability of 1, if their Q-learning

TABLE 1
Prior probability distributions

$P(\Pi_0)$		$P(I_{UEx})$		$P(RNTP)$	
Same	Diff	Yes	No	Yes	No
0.5	0.5	0.4	0.6	$High$	Low

TABLE 2
Conditional probability distributions

$P(\Pi_n^m \vert \Pi_n, RNTP)$				
Same	0	1	0	0
Diff	1	0	1	1
	Same, Yes	Same, No	Diff, Yes	Diff, No
	$\Pi_n, RNTP$			

$P(R_{UEx} \vert \Pi_n^m, I_{UEx})$				
S	0	1	1	1
F	1	0	0	0
	Same, Yes	Same, No	Diff, Yes	Diff, No
	Π_n^m, I_{UEx}			

$P(\Pi_{n+1} \vert \Pi_n, \Pi_n^m, R_{UE1}, R_{UE2})$						
Same	1	Low	Low	$High$	$f(n)$	0
Diff	0	$High$	$High$	Low	$1-f(n)$	1
	Same	Same	Same	Same	Same	Diff
	Same	Same	Same	Same	Diff	Diff
	S, S	S, F	F, S	F, F	S, S	S, S
	$\Pi_n, \Pi_n^m, R_{UE1}, R_{UE2}$					

policies are the same ($\Pi_n = Same$) and there was no RNTP exchange between the file arrivals that could change them ($RNTP = No$). In all other cases, i.e if $RNTP = Yes$ or $\Pi_n = Diff$, the masked policies of the eNBs will always be different ($Diff$). The reasoning behind the $P(R_{UEx}|\Pi_n^m, I_{UEx})$ distribution is to indicate, that a transmission to $UE1$ or $UE2$ will fail with a probability of 1 ($R_{UEx} = F$), if $I_{UEx} = Yes$ and both eNBs have chosen the same subchannel ($\Pi_n^m = Same$). If $\Pi_n^m = Diff$ or $I_{UEx} = No$, then the transmission will be successful: $R_{UEx} = S$.

The $P(\Pi_{n+1}|\Pi_n, \Pi_n^m, R_{UE1}, R_{UE2})$ table defines how the Q-learning policies of both eNBs (Π_{n+1}) are likely to change, given their current Π_n and Π_n^m, and the result of transmissions to both UEs (R_{UE1} and R_{UE2}). Firstly, if both Π_n and Π_n^m are $Same$ or both are $Diff$, and the transmissions to both UEs were successful ($R_{UE1} = R_{UE2} = S$), then both eNBs will reward their respective subchannels and maintain the same policies with a probability of 1 ($\Pi_{n+1} = \Pi_n$). Secondly, if both Π_n and Π_n^m are $Same$ and only a transmission to one of the UEs failed ($\{S, F\}$ or $\{F, S\}$), this UE is more likely to change its policy due to the WoLF learning rate used in its Q-learning algorithm, described by (6). Therefore, there is a relatively high probability of the policies being different at the next iteration: $P(\Pi_{n+1} = Diff) = High$. If transmissions to both UEs fail ($\{F, F\}$), both eNBs are likely to change their policies, thus making $\Pi_{n+1} = Same$ a more likely outcome. Lastly, if the Q-learning policies of both eNBs are the same ($\Pi_n = Same$), the masked policies are different ($\Pi_n^m = Diff$), and both transmissions are successful ($R_{UE1} = R_{UE1} = S$), the probability of the $\Pi_{n+1} = Same$ at the next iteration is time-dependent. A realistic approximation of its value at different stages of learning is:

$$f(n) = \begin{cases} 0 & n = 0 \\ 0.5 & n = 1 \\ High & n > 1 \end{cases} \quad (12)$$

If this is the first learning iteration ($n = 0$), the Q-tables of both eNBs are initialized to zeros. Therefore, if different subchannels are successfully used ($\Pi_n^m = Diff$), they will be positively reinforced and used at the next iteration with a probability of 1: $P(\Pi_{n+1} = Same|...) = 0$. After one learning iteration, there is about a 50% chance of one of the eNBs changing its Q-learning policy, depending on whether its first trial was a success on its preferred subchannel, or a failure on the other subchannel: $P(\Pi_{n+1} = Same|...) = 0.5$. Afterwards, the eNB, whose Q-learning policy was overriden by the RNTP exchange (since $\Pi_n \neq \Pi_n^m$), is relatively unlikely to change its policy due to the effect of the WoLF learning rates, i.e. the Q-values undergo smaller step changes after successful trials: $P(\Pi_{n+1} = Same|...) = High$.

The remaining 10 combinations of Π_n, Π_n^m, R_{UE1} and R_{UE2} values are not considered, since they can never occur according to the $P(\Pi_n^m|\Pi_n, RNTP)$ and $P(R_{UEx}|\Pi_n^m, I_{UEx})$ conditional probability distributions. Regardless of the values used for these combinations in the $P(\Pi_{n+1}|\Pi_n, \Pi_n^m, R_{UE1}, R_{UE2})$ table, they will be multiplied by zero during the calculation of the joint probability distribution defined in (11).

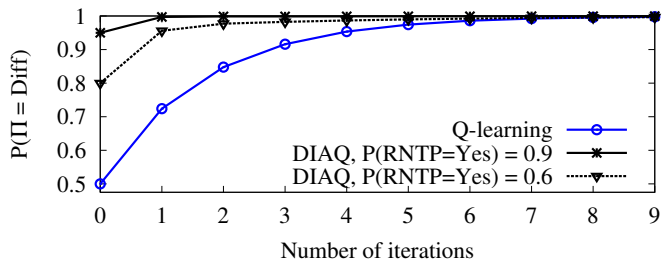
5.3 Convergence Properties

The main aim of the Bayesian network model described above is to establish the marginal likelihood of the joint Q-learning policy at the next iteration $P(\Pi_{n+1})$ by taking a sum over all other variables in P_{joint} as follows:

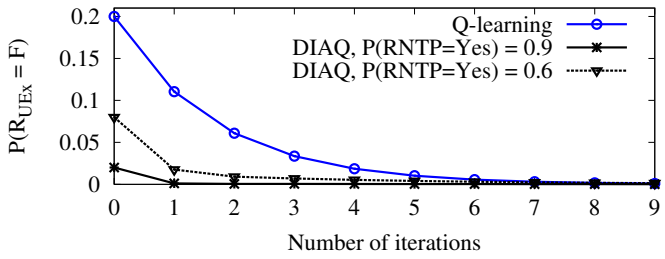
$$P(\Pi_{n+1}) = \sum_{\Pi_n} \sum_{\Pi_n^m} \sum_{R_{UE1}} \sum_{R_{UE2}} \sum_{I_{UE1}} \sum_{I_{UE2}} \sum_{RNTP} P_{joint} \quad (13)$$

The resulting distribution can then be substituted as the prior ($P(\Pi_n) \leftarrow P(\Pi_{n+1})$) for the next learning iteration. This enables iterative evaluation of the Bayesian network model which shows how the probability of transmission failure $P(R_{UEx})$ and the probability of eNBs using different subchannels $P(\Pi_n^m)$ change over time, as the learning process progresses. Both of these probability distributions can be obtained using the principle of marginalization shown in (13).

Fig. 5 shows the results of such iterative evaluation of the Bayesian network from Fig. 4. It compares the convergence performance of regular Q-learning and DIAQ with $P(RNTP = Yes)$ values of 0.9 and 0.6, respectively the cases where ICIC signalling between the neighbouring



(a) Probability of eNBs having different policies



(b) Probability of a UE being blocked or interrupted

Fig. 5. Convergence of Q-learning and DIAQ, using the probabilistic model of the 2 eNB 2 UE cellular network

eNBs is moderately reliable and relatively unreliable. The values for $High$ and Low in the conditional probability distributions in Table 2 are assumed to be $\{0.9, 0.1\}$. However, similar convergence patterns can be observed with other interpretations of "high" and "low" probabilities.

Fig. 5 demonstrates how the presence of RNTP message exchanges in DIAQ, even when they are relatively unreliable ($P(RNTP = Yes) = 0.6$), significantly speed up the learning process, especially at its early stages. The eNBs become highly likely to converge on the optimal solution ($\Pi = Diff$) significantly faster using DIAQ compared to Q-learning which only operates using trial-and-error experience. Consequently, the temporal performance of the network in terms of the probability of transmission failures shown in Fig. 5b is also superior using DIAQ.

6 Simulation Results

This section presents the results of simulating the proposed DIAQ scheme using a large scale stadium network model. The performance of this scheme is compared to that of a pure distributed Q-learning algorithm (Section 3) and a typical dynamic ICIC based scheme. The ICIC based DSA scheme assumes that each eNB always avoids transmitting on the VRBs used by its neighbours, as reported in their RNTP messages. It chooses randomly among the "safe" subchannels and blocks file transmissions when no such subchannels are available for assignment. The comparison with these two schemes is most appropriate, since they represent two key components of the DIAQ scheme separately - the RL part and the heuristic inter-eNB coordination part. The latter represents a standard approach in LTE [1][23]. Therefore, the results evaluate the importance of both of these components in the proposed DIAQ scheme.

All experiments involving ICIC signalling assume that RNTP message exchanges take place periodically every 20 ms. Therefore, the current subchannel usage of a given eNB is always mapped onto its RNTP message, since an eNB is highly likely to continue using the same subchannels for 20 ms until the next ICIC update. All eNBs are assumed to send their RNTP messages at the same time. However, the scheme will work in exactly the same way, if they are not synchronised or if the frequency of the RNTP signals is lower. Every eNB always uses the last received RNTP signal from each of its neighbours, which only affects spectrum assignment decisions for new file arrivals and does not affect current file transmissions. The RNTP threshold used for all experiments involving ICIC is -3 dB, so that the subchannels used at lower transmit powers are not included in the RNTP messages, thus increasing the potential spectrum reuse efficiency.

6.1 Stadium temporary event network

The cellular system used for simulation experiments in this paper is designed for a stadium event scenario, where a small cell LTE network is installed in a large stadium to provide an increase in mobile data capacity to the users attending the event. The network architecture is depicted in Fig. 6, where the users are located in a

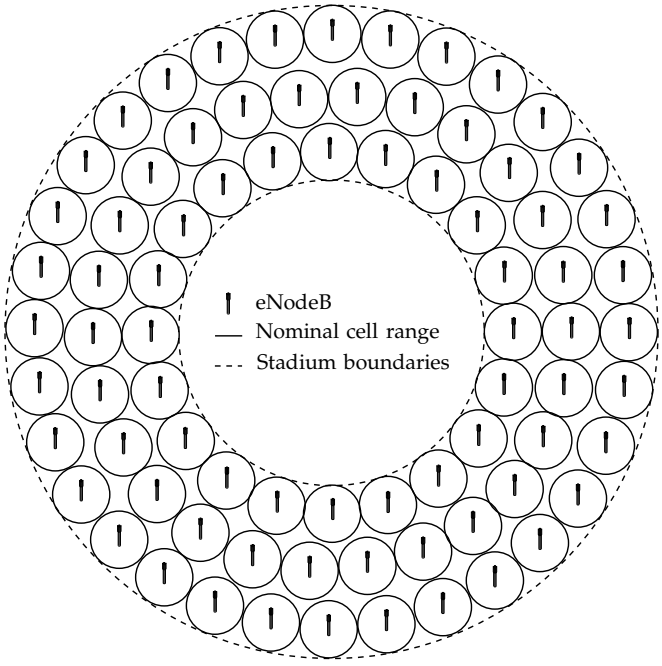


Fig. 6. Stadium network architecture

spectator area 53.7 - 113.7 m from the centre of the stadium. The spectator area is covered by 78 eNBs arranged in three rings at 1 m height, e.g. with antennas attached to the backs of the seats or to the railings between the different row levels. Seat width is assumed to be 0.5 m, and the space between rows - 1.5 m, which yields the total capacity of 43,103 seats. The parameters and assumptions used in the network model are listed in Table 3.

TABLE 3
Network model parameters and assumptions

Parameter	Value
Channel bandwidth	20 MHz: 100 LTE VRBs
Subchannel bandwidth	4 VRBs: 4 × 180 kHz [17]
Frequency band	2.6 GHz
UE receiver noise floor	94 dBm (290 K temperature, 20 MHz bandwidth, 7 dB noise figure)
Propagation model	WINNER II B3 [24]
Traffic model	3GPP FTP Traffic Model 1 [25], file size - 4.2 Mb (≈0.5 MB)
Retransmission scheduling	Uniform random back-off between 0 and 960 ms [26]
Link model	Truncated Shannon Bound model [27]
Assumptions	
Each UE is associated with an eNB with a minimum estimated downlink pathloss to it, based on the Reference Signal Received Power (RSRP)	
Open loop control of the eNB Tx power is assumed, using a constant Rx power of -74 dBm (20 dB Signal-to-Noise Ratio)	
The minimum Signal-to-Interference-plus-Noise Ratio (SINR) allowed to support data transmission is 1.8 dB [28]	

6.2 Performance Metrics

The metrics used to assess the network performance are the probability of retransmission $P(re-tx)$ and the system throughput density (STD). $P(re-tx)$ is the probability of a file transmission being blocked or interrupted, i.e. the probability of a retransmission being scheduled. It is calculated using the following equation:

$$P(re-tx) = \frac{N_{re-tx}}{N_{tx}} \tag{14}$$

where N_{re-tx} and N_{tx} are the number of retransmissions and the total number of transmissions during one sampling period respectively. STD is obtained by calculating the average system throughput during the whole simulation and dividing it by the area covered by the eNBs.

6.3 Temporal Performance

Fig. 7 compares the temporal response of the network in terms of the probability of retransmission at 1 Gbps offered traffic, using dynamic ICIC, pure distributed Q-learning and DIAQ schemes for DSA. The graph shows

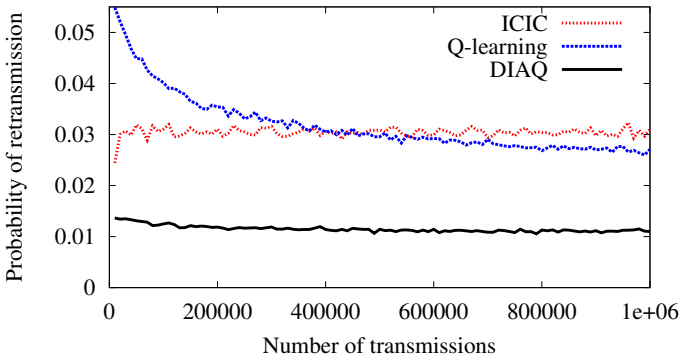


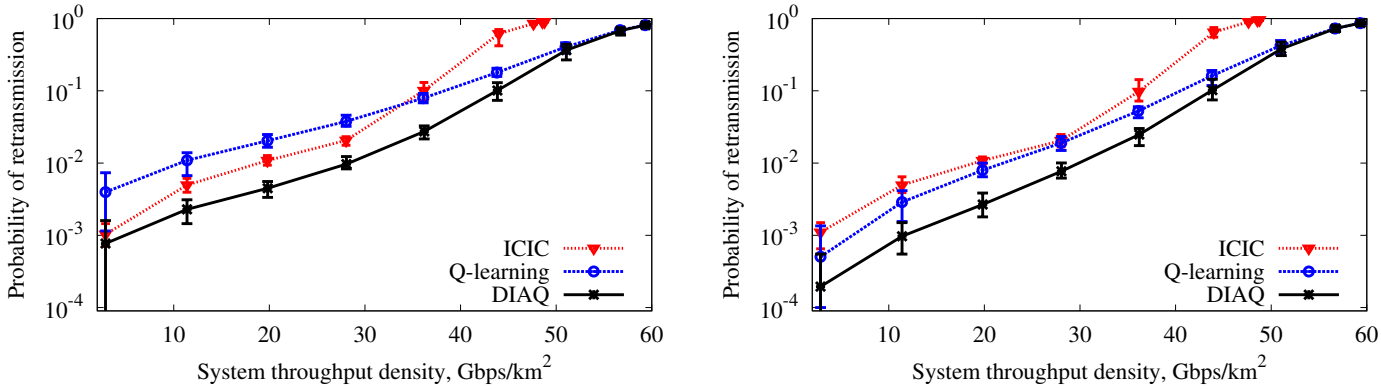
Fig. 7. Probability of retransmission time response using dynamic ICIC, pure Q-learning and distributed ICIC accelerated Q-learning (DIAQ)

the average of 50 simulations with different random seeds and UE locations in order to mitigate the noise introduced by the bursty nature of the traffic, and to produce a more statistically valid temporal response. Firstly, the graph shows that both Q-learning and DIAQ schemes converge on better DSA policies, than the ICIC scheme. Secondly, the DIAQ scheme achieves a big improvement in the initial performance compared to the classical Q-learning approach. The highly efficient guided exploration process of the DIAQ scheme results in a substantial reduction in initial $P(re-tx)$ by a factor of ≈4, compared to pure Q-learning. This improvement is consistent with the theoretically predicted outcome shown in Fig. 5b.

Fig. 7 also shows that DIAQ still has a much lower probability of retransmission compared to both schemes after 1,000,000 trials, when it is approaching its steady state. Therefore, using ICIC to enhance the Q-learning algorithm in this way dramatically speeds up its convergence, and substantially improves both its initial and steady-state performance. Such acceleration of the learning process may prove crucial in more realistic dynamically changing environments, e.g. with time-varying traffic distributions and topologies. The impact of DIAQ, compared to regular distributed Q-learning, is that it can adapt to new interference environments considerably faster. To further improve the temporal performance and robustness of RL based DSA algorithms in dynamic environments, a novel case-based RL approach is introduced by us in [19]. However, combining the case-based RL and DIAQ methods is outside of the scope of this investigation and is one of the directions for our future work.

6.4 Initial and Final Performance

Fig. 8 shows the difference in initial and final $P(re-tx)$ performance of these schemes at a wide range of traffic loads. It is plotted against the system throughput density



(a) Initial probability of retransmission

(b) Final probability of retransmission

Fig. 8. Initial and final probabilitiy of retransmission using pure ICIC, pure Q-learning and distributed ICIC accelerated Q-learning (DIAQ) at different system throughput densities

to evaluate both the QoS and the system capacity in the same graphs. The initial $P(re-tx)$ in Fig. 8a is calculated using the first 20,000 transmissions, and the final $P(re-tx)$ in Fig. 8b is calculated from the last 20,000 file transmissions. The overall simulation length is 1,000,000 file transmissions. Every data point represents the mean result of 50 different simulations at a given traffic load with the error bars showing the minimum and maximum $P(re-tx)$ in those simulations.

Fig. 8a shows that the dramatic improvement in initial performance using DIAQ instead of a classical Q-learning approach is consistent at most traffic loads. DIAQ introduces a 44-81% reduction in the initial probability of retransmission at system throughput densities below 44 Gbps/km^2. Only at ultra-high system throughput densities does the difference in their performance become negligible. DIAQ also shows a significantly better performance in initial and final probability of retransmission, compared to the dynamic ICIC scheme. Furthermore, the latter only supports system throughput densities of up to 49 Gbps/km^2, whereas DIAQ and Q-learning are significantly more robust at extremely high offered traffic densities. They both manage to support system throughput densities of up to 59 Gbps/km^2. This demonstrates that it is better to take opportunistic spectrum assignment decisions, based on reinforcement learning, instead of blocking transmissions based on ICIC signalling, since the probability of a subchannel not being occupied by any of the neighbouring eNBs tends to zero. In these cases, the heuristic ICIC approach "blindly" blocks most file transmissions, whereas Q-learning is still able to provide some insight into which subchannels could result in successful transmissions.

7 Conclusion

In this paper we propose a novel algorithm for dynamic spectrum access (DSA) in LTE cellular systems - distributed ICIC accelerated Q-learning (DIAQ). It combines distributed reinforcement learning (RL) and standardized inter-cell interference coordination (ICIC) signalling in the LTE downlink, using the framework of heuristically accelerated reinforcement learning (HARL). We also present a novel Bayesian network based approach to theoretical analysis of RL based DSA, which explains a predicted improvement in convergence behaviour achieved by DIAQ, compared to classical RL. Large scale simulation experiments of a stadium temporary event network show that it achieves superior quality of service compared to a typical heuristic ICIC approach and a state-of-the-art distributed RL based approach. It provides significantly better quality of service (QoS) in terms of the probability of retransmission and supports higher system throughput densities of up to 59 Gbps/km^2. A comparison of the probability of retransmission time response characteristics of DIAQ and pure distributed Q-learning reveals a dramatic improvement in performance at the initial stage of learning, a 44-81% improvement at all but ultra-high traffic loads, due to the use of heuristics for guiding the exploration process. This result confirms the theoretical predictions made using the Bayesian network model of the algorithm. DIAQ also exhibits excellent final performance and convergence speed. Finally, it is designed to comply with the current LTE standards. Therefore, it allows easy implementation of robust distributed machine intelligence for full self-organisation in existing commercial networks.

Acknowledgements

This work has been funded by the ABSOLUTE Project (FP7-ICT-2011-8-318632), which receives funding from the 7th Framework Programme of the European Commission.

REFERENCES

[1] S. Sesia, M. Baker, and I. Toufik, *LTE-The UMTS Long Term Evolution: From Theory to Practice*. John Wiley & Sons, 2011.

[2] G. Boudreau, J. Panicker, N. Guo, R. Chang, N. Wang, and S. Vrzic, "Interference coordination and cancellation for 4G networks," *Communications Magazine, IEEE*, vol. 47, pp. 74–81, 2009.

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

[4] T. Jiang, D. Grace, and P. D. Mitchell, "Efficient exploration in reinforcement learning-based cognitive radio spectrum sharing," *Communications, IET*, vol. 5, pp. 1309–1317, 2011.

[5] M. Bennis, S. Perlaza, P. Blasco, Z. Han, and H. Poor, "Self-organization in small cell networks: A reinforcement learning approach," *Wireless Communications, IEEE Transactions on*, vol. 12, pp. 3202–3212, 2013.

[6] X. Chen, Z. Zhao, and H. Zhang, "Stochastic power adaptation with multiagent reinforcement learning for cognitive wireless mesh networks," *Mobile Computing, IEEE Transactions on*, vol. 12, pp. 2155–2166, 2013.

[7] N. Morozs, T. Clarke, D. Grace, and Q. Zhao, "Distributed Q-learning based dynamic spectrum management in cognitive cellular systems: Choosing the right learning rate," in *IEEE International Symposium on Computers and Communications (ISCC)*, 2014.

[8] C. Watkins, "Learning from Delayed Rewards," Ph.D. dissertation, University of Cambridge, England, 1989.

[9] J. Nie and S. Haykin, "A Q-learning-based dynamic channel assignment technique for mobile communication systems," *Vehicular Technology, IEEE Transactions on*, vol. 48, pp. 1676–1687, 1999.

[10] R. Valcarce, T. Rasheed, K. Gomez, S. Kandeepan, L. Reynaud, R. Hermenier, A. Munari, M. Mohorcic, M. Smolnikar, and I. Bucaille, "Airborne base stations for emergency and temporary events," in *International Conference on Personal Satellite Services*, 2013.

[11] R. Bianchi, M. Martins, C. Ribeiro, and A. Costa, "Heuristically-accelerated multiagent reinforcement learning," *Cybernetics, IEEE Transactions on*, vol. 44, pp. 252–265, 2014.

[12] R. Bianchi and R. Lopez de Mantaras, "Case-based multiagent reinforcement learning: Cases as heuristics for selection of actions," in *European Conference on Artificial Intelligence (ECAI 2010)*, 2010.

[13] M. Simsek, M. Bennis, and A. Czylwik, "Dynamic inter-cell interference coordination in HetNets: A reinforcement learning approach," in *IEEE Global Communications Conference (GLOBECOM)*, 2012.

[14] M. Dirani and Z. Altman, "A cooperative reinforcement learning approach for inter-cell interference coordination in OFDMA cellular networks," in *International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt)*, 2010.

[15] P. Vlacheas, E. Thomatos, K. Tsagkaris, and P. Demestichas, "Autonomic downlink inter-cell interference coordination in LTE self-organizing networks," in *International Conference on Network and Services Management (CNSM)*, 2011.

[16] W. Nam, D. Bai, J. Lee, and I. Kang, "Advanced interference management for 5G cellular networks," *Communications Magazine, IEEE*, vol. 52, pp. 52–60, 2014.

[17] 3GPP, "LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures (3GPP TS 36.213 version 11.5.0 Release 11) ," Dec. 2013.

[18] C. Claus and C. Boutilier, "The dynamics of reinforcement learning in cooperative multiagent systems," in *Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence*, 1998.

[19] N. Morozs, D. Grace, and T. Clarke, "Case-based reinforcement learning for cognitive spectrum assignment in cellular networks with dynamic topologies," in *Military Communications and Information Systems Conference (MCC)*, 2013.

[20] Q. Zhao, T. Jiang, N. Morozs, D. Grace, and T. Clarke, "Transfer learning: A paradigm for dynamic spectrum and topology management in flexible architectures," in *IEEE Vehicular Technology Conference (VTC Fall)*, 2013.

[21] M. Bowling and M. Veloso, "Multiagent learning using a variable learning rate," *Artificial Intelligence*, vol. 136, pp. 215–250, 2002.

[22] T. Nielsen and F. Jensen, *Bayesian networks and decision graphs*. Springer, 2009.

[23] I. Fraimis, V. Papoutsis, and S. Kotsopoulos, "A decentralized subchannel allocation scheme with inter-cell interference coordination (ICIC) for multi-cell OFDMA systems," in *IEEE Global Telecommunications Conference (GLOBECOM)*, 2010.

[24] P. Kyösti, J. Meinilä, L. Hentilä, X. Zhao, T. Jämsä, C. Schneider, M. Narandzić, M. Milojević, A. Hong, J. Ylitalo, V. Holappa, M. Alatossava, R. Bultitude, Y. de Jong, and T. Rautiainen, "IST-4-027756 WINNER II Deliverable D1.1.2: WINNER II channel models," 2008.

[25] 3GPP, "Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Further Advancements for E-UTRA physical layer aspects (3GPP TR 36.814 version 9.0.0 Release 9)," Dec. 2010.

[26] ——, "LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol specification (3GPP TS 36.321 version 11.4.0 Release 11)," Jan. 2014.

[27] ——, "LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Frequency (RF) system scenarios (3GPP TR 36.952 version 11.0.0 Release 11)," Dec. 2012.

[28] T. Jiang, P. Li, C. Liu, N. Khan, D. Grace, A. Burr, and C. Oestges, "EU FP7 INFSO-ICT-248267 BuNGee Deliverable D4.1.2: Simulation Tool(s) and Simulation Results," 2012.

Nils Morozs (S'13) received his MEng degree in Electronic Engineering from the University of York in 2012. He is currently a PhD student in Intelligent Systems, and Communications and Signal Processing Research Groups at the Department of Electronics, University of York. His PhD research is part of the EU FP7 ABSOLUTE project concerned with developing novel LTE-compliant cognitive mechanisms for dynamic radio resource and topology management in disaster relief and temporary event networks. His research interests are in applying artificial intelligence and control engineering methods to radio resource management in cognitive cellular systems.

Tim Clarke received the B.A. degree in biology from the University of York in 1975. He joined the Royal Air Force as an Air Traffic Control Officer before becoming an Education Officer. He underwent advanced training at the Royal Military College of Science, Shrivenham, where he received the M.Sc. degree in guided weapons systems engineering. He is Senior Lecturer in Control Engineering and is Head of the Control Systems Laboratory, Intelligent Systems Group, Department of Electronics, University of York. His research interests are in the areas of biologically inspired engineering and control systems.

Mr. Clarke is a member of the IET and serves on IFAC Technical Committees 5.4 (Large Scale Complex Systems), 7.3 (Aerospace), and 7.5 (Intelligent Autonomous Vehicles).

David Grace (S'95-A'99-M'00-SM'13) received his PhD from University of York in 1999, with the subject of his thesis being Distributed Dynamic Channel Assignment for the Wireless Environment. Since 1994 he has been a member of the Department of Electronics at York, where he is now a Senior Research Fellow and Head of Communications and Signal Processing Research Group. He is also a Co-Director of the York - Zhejiang Lab on Cognitive Radio and Green Communications, and a Guest Professor at Zhejiang University. Current research interests include cognitive green radio, particularly applying distributed artificial intelligence to resource and topology management to improve overall energy efficiency; architectures for beyond 4G wireless networks; dynamic spectrum access and interference management. He is a one of the lead investigators on FP7 ABSOLUTE which is dealing with extending LTE-A for emergency/temporary events through application of cognitive techniques, and recently a co-investigator of the FP7 BuNGee project dealing with broadband next generation access. He is an author of over 180 papers, and author/editor of 2 books. He currently chairs IEEE Technical Committee on Cognitive Networks and the Worldwide Universities Network Cognitive Communications Consortium (WUN CogCom), and is a member of COST IC0902. He is a founding member of the IEEE Technical Committee on Green Communications and Computing. In 2000, he jointly founded SkyLARC technologies Ltd, and was one of its directors.