Learning-based Framework for Policy-aware Cognitive Radio Emergency Networking

Eun Kyung Lee, Hariharasudhan Viswanathan, and Dario Pompili
Department of Electrical and Computer Engineering, Rutgers University, New Brunswick, NJ, USA
E-mails: {eunkyung_lee, hari_viswanathan, pompili}@cac.rutgers.edu

***Abstract*—Uncertainties in the wireless communication medium do not allow for guarantees in network performance for cognitive radio applications envisaged for mobile ad hoc emergency networking. The novel concept of mission policies, which specify the Quality of Service (QoS) requirements of the incumbent network as well as the cognitive radio networks, is introduced. The use of *mission policies*, which vary over time and space, enables graceful degradation in the QoS of incumbent network (only when necessary) based on mission-policy specifications. A Multi-Agent Reinforcement Learning (MARL)-based cross-layer communication framework, *RescueNet*, is proposed for self-adaptation of nodes in cognitive radio networks. Also, the novel idea of *knowledge sharing* among the agents (nodes) is introduced to significantly improve the performance of the proposed solution.**

***Index Terms*—Cognitive Radio, Licensed Spectrum, Mission Policies, Reinforcement Learning, Multi-agent Systems.**

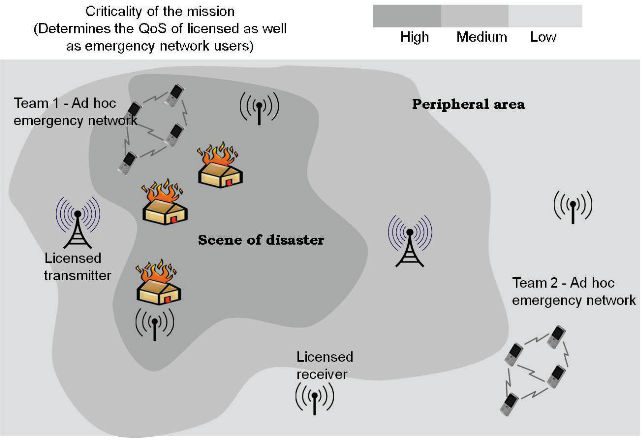

Criticality of the mission
(Determines the QoS of licensed as well as emergency network users)
High
Medium
Low
Team 1 - Ad hoc emergency network
Peripheral area
Scene of disaster
Licensed transmitter
Team 2 - Ad hoc emergency network
Licensed receiver

Fig. 1. Ad hoc emergency networks operating in the vicinity of licensed incumbents in the event of an emergency. Mission policies, which reflect the criticality and, hence, the QoS of both networks, *vary over space* (depending on proximity to the scene of the disaster) and *over time* (depending on the phase of the mission).

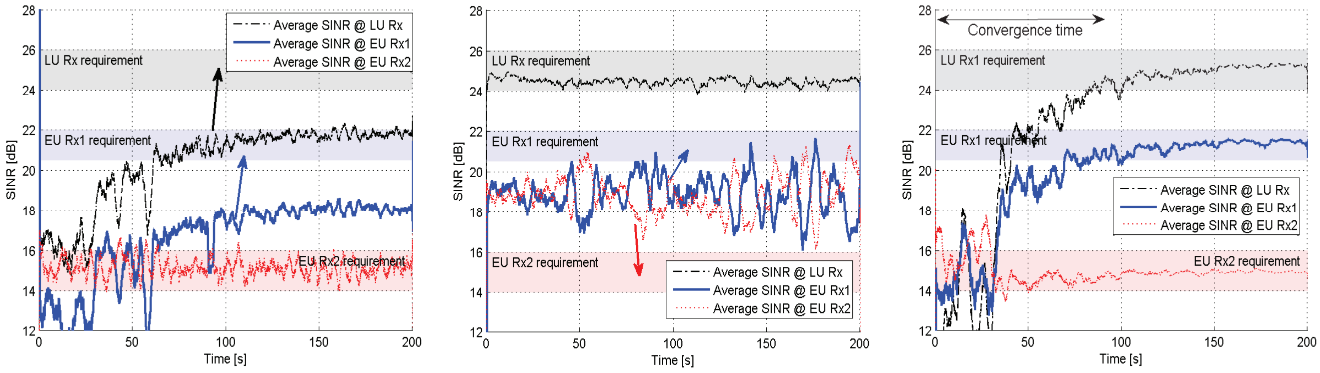
I. Introduction

Reliable and high data-rate wireless multimedia communication (e.g., images, voice, and live video streams) among mobile computing devices is becoming a fundamental requirement for *emerging* wireless applications such as emergency networking, smart-grid, and body-area networking. The impracticality of dedicating spectrum resources for each futuristic wireless application has led to the emergence of Cognitive Radio Networking (CRN) in licensed spectrum as the most promising wireless networking paradigm of the future.

However, the use of various non-interoperable CRN solutions by different applications prevents seamless information sharing among Cognitive Radio (CR) nodes of different networks and does not guarantee any form of Quality of Service (QoS) to both the licensed incumbent network and the co-existing CR networks. In an effort to provide such statistical guarantees, the wireless networking research community tried to analyze uncertainties in wireless environment by modeling or controlling their causes. Moreover, uncertainties (or *non-stationarity*) in the wireless communication medium – due to its shared nature (limited bandwidth), time-varying characteristics, network attacks, and node mobility – do not always allow for guarantees in terms of reliability and network performance. Uncertainties render conformance with specified QoS requirements at all the nodes in the wireless network a significant challenge. The causes for uncertainties are hard to model because the wireless environment changes over time and space based on the choice of many network parameters associated with different protocol layers at various nodes – transmission power, modulation, and error correction in the *physical layer*; medium access parameters in the *link layer*; routing scheme in the *network layer*; and traffic pattern in the *application layer*. Note that the network nodes may vary their network parameters either in response to changes in their immediate environment or in response to changes in the high-level application QoS requirements (also referred to as "policies") in time and space.

An "optimal" choice of parameters may be obtained by solving a centralized cross-layer networking optimization problem, which optimizes network parameters in different layers, based on unrealistic assumptions such as instantaneous knowledge of global network state, complete knowledge of incumbent user performance, and availability of infinite computational capabilities. Another approach is solving a number of localized optimization problems (based only on locally observed and shared information), which cannot balance the opposing requirements of capturing local interference constraints as well as satisfying end-to-end (e2e) QoS requirements of the CRN and incumbent network traffic. To eliminate reliance on unrealistic assumptions, we propose a *model-free* communication framework based on Multi-Agent Reinforcement Learning (MARL) [1] for *self-adaptation of coordinating CR nodes (autonomous agents).* Our distributed solution *converges* to a local *optimal joint control policy* (i.e., optimal choice of transmission param-

978-1-4799-1353-4/13/$31.00 ©2013 IEEE



(a) (b) (c)

Fig. 2. (a) Average Signal to Interference-plus-Noise Ratio (SINR) at EU and LU receivers when EU transmitters perform transmission power control using *Independent MARL* (only EU Rx2's SINR requirement is met); (b) Average SINR at EU and LU receivers when EU transmitters perform transmission power control using *Global reward MARL* (only LU Rx's SINR requirement is met); (c) Average SINR at EU and LU receivers when EU transmitters perform transmission power control using our hybrid approach based on *DVF-MARL* (LU Rx's SINR requirement as well as the EU Rxs' requirements are met).

eters at all agents in the neighborhood) through coordination. The optimal control policy ensures *conformance* to the QoS requirements of the CR and incumbent networks as specified by the high-level policy. We also address the challenges to the convergence of our MARL-based approach posed by the non-stationarity of the environment in our problem (due to dynamic mission policies and node mobility) on the fly.

The proposed concepts are shown in Fig. 1 using ad hoc emergency networking. Emergency networks deployed on rescue and recovery missions in the aftermath of disasters are usually composed of multiple teams of first responders, or Emergency Users (EUs), dispersed over a wide geographical region coexisting with Licensed Users (LUs). The data and voice traffic, and the corresponding QoS requirements of EUs as well as LUs, are tightly coupled to the criticality of the mission. In this paper, we present a policy-aware MARL-based communication framework for ad hoc emergency networking, called RescueNet. There have been prior efforts in applying distributed multi-agent reinforcement learning for wireless networking [2] as well as specific studies on RL for spectrum sensing, scheduling, and selection [3] in mesh networks, QoS support in wireless sensor networks with and without relay selection [4], [5], and sensing coverage [6].

Presently, to the best of our knowledge, there are no *MARL-based policy-aware* CR solutions. Also, our solution represents a shift from the established primary-secondary model, which uses fixed priorities, towards graceful QoS change of LUs (only when necessary) based on policies. The following are our contributions: i) we cast the ad hoc CR emergency networking problem as a MARL problem, i.e., identify states, actions and rewards, and design a flexible reward function that captures the degree of conformance to the high-level policies; ii) we address the significant challenge to the convergence of the learning process posed by the non-stationarity of the problem of ad hoc CR networking in licensed spectrum; iii) we introduce the novel idea of *transferring knowledge* from experienced agents to young agents to expedite the learning process, and, hence, the conformance to policies.

The rest of this paper is organized as follows: in Sect. II, we provide the necessary background on RL, motivate the need for our framework, and present the RescueNet framework; in Sect. III, we evaluate the performance of RescueNet in terms of convergence and conformance; finally, in Sect. IV, we draw our conclusion.

II. Proposed Solution

A. Background and Motivation

We provide here the necessary background on RL and then present the intuitions and preliminary analysis behind our choice of a *hybrid learning strategy* (distributed yet localized) as the foundation of our policy-aware CR networking framework.

Multi-Agent Reinforcement Learning: Policy-aware CR networking in licensed spectrum resembles a multi-agent system trying to converge to the optimal joint control policy in a *distributed* manner. The generalization of the multi-agent system is the Markov Decision Process (MDP) specified by the following tuple: $\langle \boldsymbol{S}, \boldsymbol{A}, \phi, \rho \rangle$, where the discrete sets of environment states, $\boldsymbol{S} = \prod_{i \in \mathcal{M}} \mathcal{S}_i$, and actions, $\boldsymbol{A} = \prod_{i \in \mathcal{M}} \mathcal{A}_i$ are made up of individual agent states and actions. Here, $\mathcal{M}$ represents the set of autonomous agents in the multi-agent system. It is important to note that the transition function $\phi()$ and reward function $\rho()$ depend on the *joint* environment state and action information, which is not available at any individual agent. Hence, ensuring convergence in a multi-agent scenario requires coordination among the autonomous agents. There are three possible approaches to solve MARL problems using *Q-learning*, a model-free RL technique. The state-action pair's goodness value is called the *Q-value* and the function that determines the Q-value is called the *Q-function*. An agent can find an optimal control policy by approximating iteratively its Q-values using prior estimates, short-term reward $r = \rho(s, a) \in \mathbb{R}$, where $s \in \boldsymbol{S}$ and $a \in \boldsymbol{A}$, and discounted future reward. We explain each of those approaches with an example of emergency-networking scenario and motivate the need for our hybrid approach, which is explained in Sect. II-B.

Toy example scenario: The transmitters (EU Tx1 and EU Tx2) of two EU pairs operating in the vicinity of a LU pair perform transmission power control to ensure that the Signal to Interference-plus-Noise Ratios (SINRs) at their receivers (EU Rx1 and EU Rx2) are within prescribed intervals, which are depicted as shaded regions in Fig. 2(a-c). *These SINR requirements at both the emergency and the incumbent network nodes represent a simple mission-policy specification.* The different SINR requirements are derived directly from the corresponding throughput requirements as SINR dictates the achievable channel efficiency in $\mathrm{bps/Hz}$. All the devices operate in the same frequency band and the EU transmitters choose from one of the possible five power levels (4 to 20 dBm in steps of 4 dB). Log-distance path loss model is used to calculate the transmission loss.

Independent MARL [7]: In this fully distributed approach, each agent acts independently *without* coordination. The Q-learning procedure at a node can be summarized as,

$$Q_{n+1}(s,a) = (1-\alpha_n)Q_n(s,a) + \alpha_n \big[r + \gamma \max_{a' \in \mathcal{A}} Q_n(s',a')\big],$$

where $\alpha_n \in (0,1]$ is the *learning factor* and $\gamma \in [0,1)$ is the *discount factor*. Mission-policy conformance in emergency networking depends heavily on intra-emergency-network and inter-network (emergency and incumbent) interference. As independent MARL does not allow for any information exchange among the agents, it is impossible to mitigate the intra-emergency-network interference and, hence, there is no guarantee for conformance even to simple one-sided mission policies that do not guarantee any QoS to the LUs. Figure 2(a) shows the average SINR at the EU and LU receivers when the EUs try to satisfy their own QoS requirements without any coordination and feedback about the LUs' performance.

Global reward MARL [8]: In this approach, even though the agents are only aware of their individual states and actions (exactly as in Independent MARL), the Q-value estimates are updated based on a global reward that is disseminated across all the agents. The aggregated interference generated by the emergency network nodes at the incumbent users can be measured and a global reward can be estimated based on the QoS experienced by the LUs. However, the intra-emergency-network dynamics (effect of joint actions at each EU Rx) cannot be captured at a central entity. Hence, global reward MARL can only support mission policies that convey the QoS of the LUs alone, making it unsuitable for emergency networking. The average received SINR at the EU and LU receivers when Global reward MARL is employed by the emergency network nodes is shown in Fig. 2(b).

Distributed Value Function (DVF) MARL [9]: In this approach, the Q-value estimates at each autonomous agent are updated based on the individual short-term rewards as well as on additional information obtained from other agents in the neighborhood. Neighborhood here refers to a group of agents that are within the radio communication range of each other. Every agent exchanges *the largest* Q-value that is associated with its current state with every other agent in its neighborhood. More complex strategies taking into account the fact that not all neighbors are equally affected by the actions of an agent are possible. The additional information obtained from agents ensures that the agent takes into account the effect of its own actions on all its neighbors. DVF-MARL approach can support mission policies that convey the QoS requirements of the emergency networks due to its ability to capture in-network dynamics. However, it cannot support a two-sided mission policy (which specifies both EU and LU QoS) due to the inability to capture the effect of EUs' actions on the LUs.

B. An Hybrid Approach: our RescueNet Framework

In order to effectively support two-sided mission policies, we propose a hybrid learning approach that incorporates localized feedback (either partial or full) regarding the effect of its own actions on the neighboring EUs *as well as* the LUs. The performance of such an approach is shown in Fig. 2(c). The convergence of the hybrid approach exhibits sensitivity to initial states and to the choice of the three learning parameters, namely, *exploration factor* ϵ, *learning factor* α, and *discount factor* γ. Longer convergence times may hamper critical communication among the EUs. Moreover, conformance to the specified mission policy is determined by how well the reward function captures the dynamics between the e2e behavior and the effect of an agent's action on its neighborhood (observed through information exchange).

States: We represent the state of each network node $s^i \in \mathcal{S}^i$ as a tuple $\langle F^i_{min}, BW^i, \eta^i, P^i, M^i, R^i, k \rangle$, where the starting frequency F^i_{min} [Hz] and bandwidth BW^i [Hz] together represent the frequency band of operation, η^i represents the modulation and coding scheme, P^i [W] is the transmission power, M^i and R^i are parameters associated with the Medium Access Control (MAC) and routing layers, and, finally, k is the destination node to which node i is currently sending data packets. Note that M^i may correspond to a specific time slot, random access delay, or spreading factor depending on the type of MAC used; and R^i corresponds to the specific routing protocol employed to select the next hop.

Reward function: The reward function uses direct feedback from the environment and the QoS requirements specified by the mission policy to produce scalar rewards whose magnitude conveys the degree of conformance with the high-level policy. The reward function produces an aggregated reward $r^{i,tot}$ at agent i (source) by incorporating feedback from agent k (destination) about e2e goodput (gp^{ik}) and delay (d^{ik}) as well as SINR about the incumbent network performance (lu). The positive reward for delay performance is high (i.e., close to the maximum reward value of 1) when the achieved average delay is close to the minimum delay requirement. The positive reward for goodput performance is high if the achieved goodput is close to the maximum goodput requirement. This specific choice of positive reward values indicates a preference towards short transmission times so to minimize packet collisions and costly retransmissions. The agents receive negative rewards (or penalties) if they do not conform to the mission policy's requirements. The magnitude of the rewards (in conjunction with the learning and discount factors) have been chosen to ensure that the Q-value estimates do not fluctuate drastically with a single reward. The mission policy specifies the QoS requirements of the emergency network in terms of minimum and maximum values. The reward function uses these values to give scaled positive rewards when the requirements are met and to give negative rewards when they are not.

Exploration-exploitation trade-off: *Non-stationarity* of the environment in ad hoc emergency networks can be attributed to time-varying mission policies, dynamics of the emergency and incumbent network traffic, node mobility, and the time-varying wireless channel. We overcome the challenge for the stabilization of the learning procedure by balancing exploration-exploitation trade-off and by appropriately choosing the learning factor. The exploration factor ϵ (of the ϵ-greedy approach) is time varying with a very high value in the beginning of each static game (*more exploration*) and with very low value at the end of each static game (*more exploitation*). We determine the decay rate δ_ϵ of the exploration factor at all agents based on the degree of mobility, i.e., $\delta_\epsilon = \psi(v)$, where v is the average speed of all nodes in the emergency network. In case of low mobility, nodes should exploit their knowledge more as their environment changes very slowly. In higher mobility, nodes should explore more than they exploit as their acquired knowledge may become outdated sooner. The evolution of the exploration factor over time is given by $\epsilon^i_{n+1} = \epsilon^i_n \cdot \delta_\epsilon$.

Knowledge sharing among agents: RescueNet can enable convergence of multiple agents to an optimal joint control policy, but the convergence takes time as the process of Q-learning requires exploration of all possible local control policies with non-zero probability. When the mission policy changes over time, the agents have to learn the new optimal joint control policy all over again. To expedite the convergence, we propose a novel mechanism of *knowledge sharing* among agents. This bootstraps the new agents to start from a good initial state as well as to use a significantly higher exploitation rate and a significantly lower learning rate than the usual so that they can converge to an optimal joint control policy much faster (Fig. 3) than they would have under usual circumstances (Fig. 2(c))

Specification of learning factor: The learning factor determines the weights associated with prior experience and with the new information in the iterative approximation of the Q-function. In RescueNet, the learning factor is time varying in order to ensure stabilization of the learning process, i.e., greater importance is given to new information initially in the static game while prior experience is leveraged more as time progresses. The decay rate δ_α of learning factor at all agents depends not only on the stage of the static game but also on the degree of node mobility, i.e., $\delta_\alpha = \sigma(v)$. In the case of very high node mobility, nodes should refrain from using their experience as it may be outdated. The time evolution of the learning factor is given by $\alpha^i_{n+1} = \alpha^i_n \cdot \delta_\alpha$.

III. Performance Evaluation

To evaluate the performance of RescueNet, we implemented it on ns3, a packet-based discrete-event network simulator.

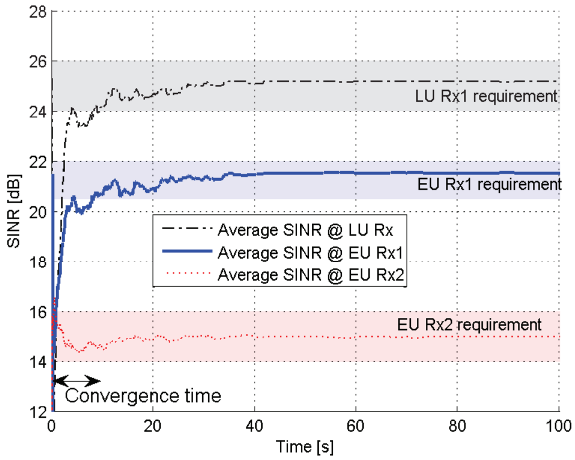


Fig. 3. Average SINR at EU and LU receivers when EU transmitters perform transmission power control using our hybrid approach along with *knowledge sharing*, i.e., sharing the knowledge of good initial states for the learning process (both LU and EU Rx's SINR requirements are met).

We performed two different types of simulations: in *static* and *mobility* scenarios. The tunable transmission parameters and assumptions regarding the loss model, MAC, and routing schemes are listed in Table I.

TABLE I
SIMULATION SETTINGS AND TUNABLE TRANSMISSION PARAMETERS

Transmission power	$4 - 20$ dBm in steps of 4 dB
Transmission band	3 channels in 515-527 MHz band (4MHz wide)
Modulation scheme	8-, 16-, 32-QAM
MAC	DS-CDMA with chaotic spreading codes [10]
Routing	Most Forward within Radius (MFR) [11]
Loss model	Log-distance path loss model

Static scenario: The topology of EU and LU nodes used in this scenario is depicted in Fig. 4, which shows two teams of EUs operating in the vicinity of a LU receiver.

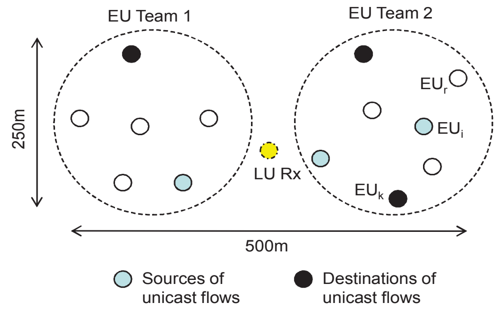
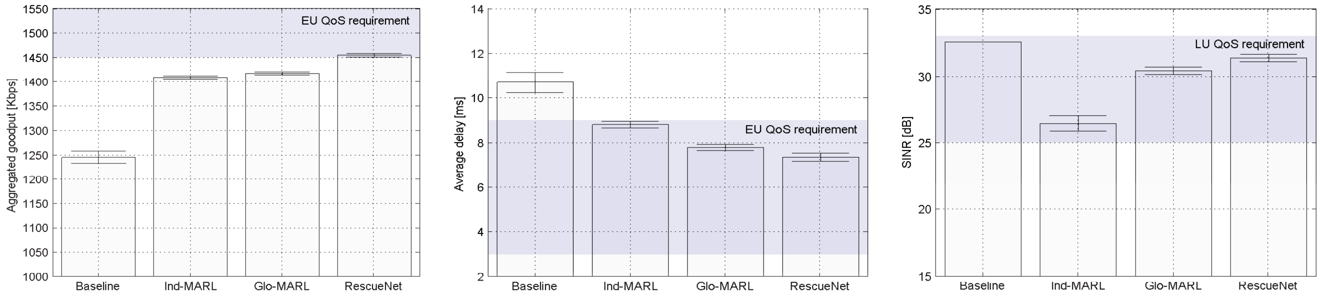


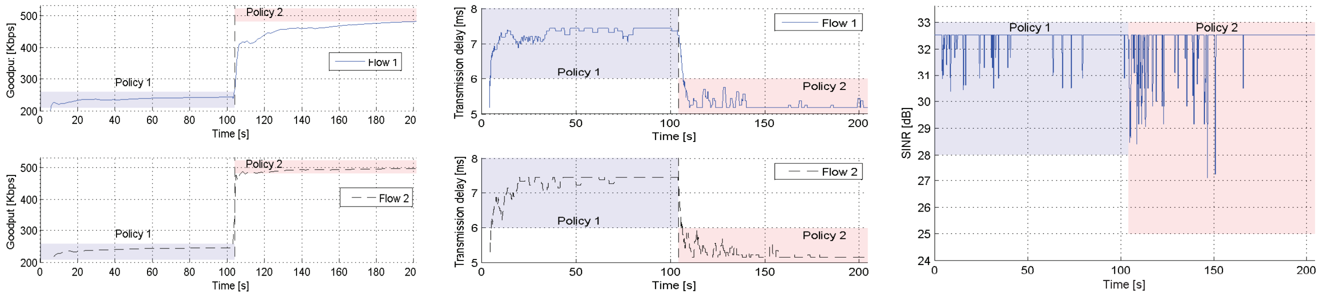
Fig. 4. Static scenario used to evaluate the performance of RescueNet in terms of *conformance* to mission policy and *convergence* to an optimal control policy. Two teams of EU nodes operating in the vicinity of a LU.

We compared RescueNet with i) a framework that employs the localized optimization approach similar to the ones proposed in [12], [13] (referred to as "Baseline"), ii) a fully distributed independent MARL-based framework (referred to as "Ind-MARL"), and iii) a global reward MARL-based framework (referred to as "Glo-MARL"). The EUs decide on the



(a) (b) (c)

Fig. 5. Static Scenario: Emergency network performance in terms of (a) aggregated goodput, (b) average packet delay, and (c) average SINR at LU receivers when EU nodes employ a local optimization approach, independent MARL (Ind-MARL), global reward MARL (Glo-MARL), and RescueNet.



(a) (b) (c)

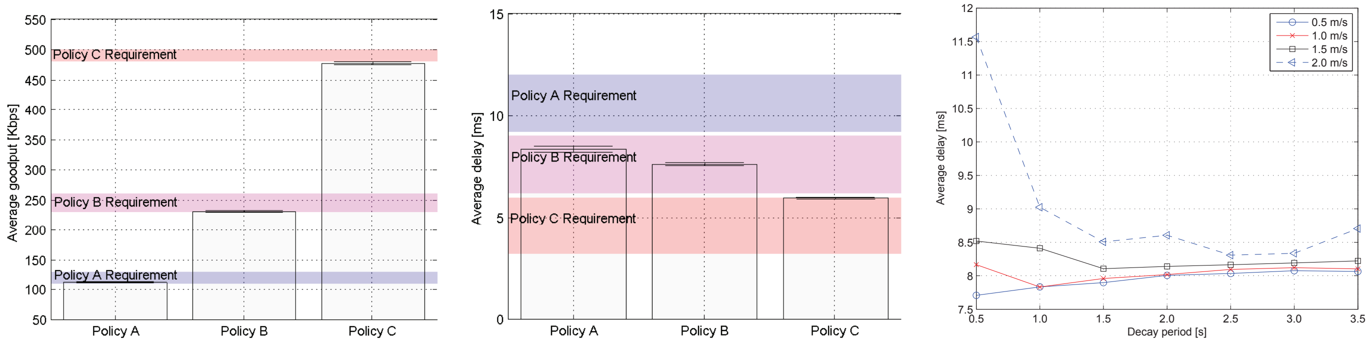
Fig. 6. Static Scenario: RescueNet's ability to adapt to time-varying mission policies (a) Goodput of flows 1 and 2; (b) Average packet delay of flows 1 and 2; and (c) Average SINR at the LU receiver when the mission-policy specification changes over time.

appropriate values of the following transmission parameters in a cross-layer manner: transmission power level, frequency band of operation (from 2 channels), and modulation scheme using one of the four aforementioned frameworks.

The data traffic in the EU teams was assumed to be three unicast flows, 500 Kbps each, one in Team 1 and two in Team 2. Figures 5(a) and (b) show the aggregated goodput and average packet delay, respectively, of all the three unicast flows in the emergency network. Figure 5(c) shows the average SINR measured at the LU. It can be observed that the emergency network fully conforms to the mission-policy specification (depicted as blue-shaded regions) when it employs RescueNet. However, the policy is violated when the other three frameworks are employed for self-adaptation. EUs employing Ind-MARL try to satisfy only the QoS requirements of the flows that they handle and do not consider the effect of their actions on both the neighboring EUs and the incumbent network nodes. As a result, the EU unicast flows suffer from high delays due to packet collisions, which also affects their goodput. The performance of incumbents is also adversely affected and that is evident from the average SINR measurements at the LU receiver. When EUs employ Glo-MARL, the incumbent network performance is guaranteed, as shown in Fig. 5(c). However, the EUs do not account for their own QoS and, hence, violate the pre-specified mission-policy specifications. The localized optimization approach, Baseline, suffers the most in terms of performance because of its inability to capture global network dynamics (it relies only on local observations). To account for the effect of an agent's action on its neighbors, Baseline needs information about ongoing receptions, the received power, and the noise interference levels in each frequency channel. Hence, the Baseline incurs an overhead to exchange such information while not guaranteeing any optimality. Conversely, in RescueNet, agents in the vicinity coordinate to tackle intra-emergency-network interference by exchanging only the maximum state-action-pair values associated with their current states. Hence, EUs not only conform to their own QoS requirements but also take into account the effect of their actions on their neighbors.

One of the main attributes of RescueNet is its ability to conform to dynamic time-varying mission policies. To verify this ability, we used the set up depicted in Fig. 4 but with only Team 2 operating in the vicinity of a LU receiver. The EUs decide on the appropriate values of the following transmission parameters in a cross-layer manner: transmission power level and modulation scheme using the RescueNet framework. Figures 6(a) and (b) show the average (moving window) goodput and packet delay, respectively, of each unicast flow in the team of EUs. Figure 6(c) shows the average SINR measured at the LU. It can be observed that the emergency network employing the RescueNet framework fully conforms to the time-varying QoS requirements (shaded regions). This flexibility is due to the generic nature of the reward function of the proposed RescueNet framework.

Mobility scenario: To obtain results that will show conclusively RescueNet's ability to adapt to non-stationarity in the operating environment, we performed simulations with node mobility as well as with time- and space-varying mission



(a) (b) (c)

Fig. 7. Mobility Scenario: (a) Average goodput and (b) delay of flows corresponding to policies A, B, and C when 3 mobile teams under 3 distinct policies are operating in the vicinity of a LU pair. Average node speed in this scenario is $1\ \mathrm{m/s}$; (c) Impact of the periodicity of decay of learning factor and of growth of exploitation factor on conformance with a mission policy at different node velocities.

policies. The EUs in all teams perform a random walk (for a randomly chosen duration between 5 and 20 s) within a rectangular area ($200 \times 200\ \mathrm{m}^2$) around their initial positions with a pause (randomly chosen between 5 and 10 s) after every walk. This mobility pattern simulates movement patterns of first responders in the scene of disaster by incorporating uniformly distributed random walking durations and random pause durations. This was done to eliminate any bias that may be introduced by a static network topology.

Our simulation results in Figs. 7(a) and (b) show clearly that the average goodput and average delay of unicast flows corresponding to the three different mission policies A, B, and C are very close to the goodput and delay specifications of each of those policies with small relative confidence intervals. The average SINR at the LU receiver did not drop below the minimum required SINR (25 dB) at any point in time during the experiments. The consistency in the performance of RescueNet under node mobility clearly demonstrates its ability to adapt to dynamic time- and space-varying mission policies as well as to non-stationarity in the environment. Figure 7(c) compares the performance of the emergency network in terms of average packet delay at various decay periodicity values for four different average node velocities. We can observe that, as the node velocity increases, the decay period has to be increased to achieve delays that conform to the mission policy. This is due to the fact that, at higher node velocities, the knowledge acquired by agents do not hold for long and, hence, they should have the capability to learn and adapt to the new environment quickly.

IV. Conclusion

We envisioned a policy- and learning-based paradigm for Cognitive Radio (CR) networking in licensed spectrum. We introduced the concept of mission policies, which specify the Quality of Service (QoS) for CR as well as incumbent network traffic. The learning-based paradigm for CR networks enables graceful QoS change of incumbent networks based on mission-policy specifications. We developed a Multi-Agent Reinforcement Learning (MARL)-based communication framework, RescueNet, for realizing this new paradigm and showed fast conformance to dynamic time-varying mission policies.

References

[1] L. Busoniu, R. Babuska, and B. De Schutter, “A Comprehensive Survey of Multiagent Reinforcement Learning,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 38, no. 2, pp. 156 –172, Mar. 2008.

[2] C.-K. Tham and J.-C. Renaud, “Multi-Agent Systems on Sensor Networks: A Distributed Reinforcement Learning Approach,” in *Proc. of Intelligent Sensors, Sensor Networks and Information Processing Conference (ISSNIP)*, Dec. 2005.

[3] M. Di Felice, K. Chowdhury, A. Kassler, and L. Bononi, “Adaptive Sensing Scheduling and Spectrum Selection in Cognitive Wireless Mesh Networks,” in *Proc. of Computer Communications and Networks (ICCCN)*, Aug. 2011.

[4] X. Liang, M. Chen, Y. Xiao, I. Balasingham, and V. Leung, “A Novel Cooperative Communication Protocol for QoS Provisioning in Wireless Sensor Networks,” in *Proc. of Testbeds and Research Infrastructures for the Development of Networks Communities and Workshops (TridentCom)*, Apr. 2009.

[5] X. Liang, I. Balasingham, and V. Leung, “Cooperative Communications with Relay Selection for QoS Provisioning in Wireless Sensor Networks,” in *Proc. of Global Telecommunications Conference (GLOBECOM)*, Dec. 2009.

[6] J.-C. Renaud and C.-K. Tham, “Coordinated Sensing Coverage in Sensor Networks using Distributed Reinforcement Learning,” in *Proc of IEEE International Conference on Networks (ICON)*, Sep. 2006.

[7] M. L. Littman, “Value-function Reinforcement Learning in Markov Games,” *Cognitive Systems Research*, vol. 2, no. 1, pp. 55 – 66, Apr. 2001.

[8] A. Galindo-Serrano and L. Giupponi, “Distributed Q-Learning for Aggregated Interference Control in Cognitive Radio Networks,” *IEEE Transactions on Vehicular Technology*, vol. 59, no. 4, pp. 1823 –1834, May 2010.

[9] J. G. Schneider, W. Wong, A. W. Moore, and M. A. Riedmiller, “Distributed Value Functions,” in *Proc. of International Conference on Machine Learning (ICML)*, Bled, Slovenia, Jun. 1999.

[10] D. Broomhead, J. Huke, and M. Muldoon, “Codes for Spread Spectrum Applications Generated Using Chaotic Dynamical System,” *Dynamics and Stability of Systems*, vol. 14, no. 1, pp. 95–105, Mar. 1999.

[11] H. Takagi and L. Kleinrock, “Optimal transmission ranges for randomly distributed packet radio terminals,” *IEEE Transactions on Communications*, vol. 32, no. 3, pp. 246–257, Aug. 1984.

[12] L. Ding, T. Melodia, S. Batalama, J. Matyjas, and M. Medley, “Cross-Layer Routing and Dynamic Spectrum Allocation in Cognitive Radio Ad Hoc Networks,” *IEEE Transactions on Vehicular Technology*, vol. 59, no. 4, pp. 1969 –1979, May 2010.

[13] D. Pompili and I. F. Akyildiz, “A Multimedia Cross-Layer Protocol for Underwater Acoustic Sensor Networks,” *IEEE Transactions on Wireless Communications*, vol. 9, no. 9, pp. 2924–2933, Jul. 2010.