Computer Networks 98 (2016) 14–28

Contents lists available at ScienceDirect

Computer Networks

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

RescueNet: Reinforcement-learning-based communication framework for emergency networking

Eun Kyung Lee*, Hariharasudhan Viswanathan, Dario Pompili

NSF Center for Cloud and Autonomic Computing, Department of Electrical and Computer Engineering, Rutgers University-New Brunswick, NJ, USA

ARTICLE INFO

Article history:
Received 6 May 2014
Revised 17 December 2015
Accepted 13 January 2016
Available online 4 February 2016

Keywords:
Mission policies
Reinforcement learning
Multi-agent systems
Licensed spectrum
Emergency networking

ABSTRACT

A new paradigm for emergency networking is envisioned to enable reliable and high data-rate wireless multimedia communication among public safety agencies in licensed spectrum while causing only acceptable levels of disruption to incumbent network communication. The novel concept of *mission policies*, which specify the Quality of Service (QoS) requirements of the incumbent networks as well as of the emergency networks involved in rescue and recovery missions, is introduced. The use of mission policies, which vary over time and space, enables *graceful degradation* in the QoS of the incumbent networks (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 emergency networks based on this new paradigm. In addition to addressing the research challenges posed by the non-stationarity of the problem, the novel idea of *knowledge sharing* among the agents of different ages (either bootstrapping or selective exploration strategies or both) is introduced to improve significantly the performance of the proposed solution in terms of convergence time and conformance to the mission policies.

1. Introduction

Reliable and high data-rate wireless multimedia communication (e.g., images, voice, and live video streams) among public safety agencies is a fundamental requirement for efficient *rescue and recovery* missions in the aftermath of natural (e.g., earthquakes, hurricanes) and man-made disasters (e.g., terrorist attacks, industrial accidents). However, the use of various non-interoperable communication technologies (e.g., terrestrial trunked radio, analog radio networks, GSM, UMTS, LTE) by different national and international agencies prevents seamless information sharing among different teams of first responders [1], Conditional Auction, law enforcement groups, hospitals, military personnel, and among rescue shelters [2]. Also, the responding agencies cannot depend on existing wireless infrastructure networks for interoperability as such infrastructure may have failed or be oversubscribed during emergencies.

Allocation of dedicated spectrum was recently considered as a possible solution [3] for a seamless and fully interoperable emergency networking system in the US. However, dedicated spectrum may increase the network vulnerability to jamming attacks, lead to heavy under-utilization of scarce spectrum resources during non-emergency periods, and suffer from the problem of over-subscription during catastrophic events. Hence, some parties have favored a plan whereby airwaves could be "conditionally auctioned" off to commercial wireless carriers (possibly at

* Corresponding author. Tel.: +8325451949.
E-mail addresses: eunkyung_lee@cac.rutgers.edu (E.K. Lee), hari_viswanathan@cac.rutgers.edu (H. Viswanathan), pompili@cac.rutgers.edu (D. Pompili).

http://dx.doi.org/10.1016/j.comnet.2016.01.011
1389-1286/

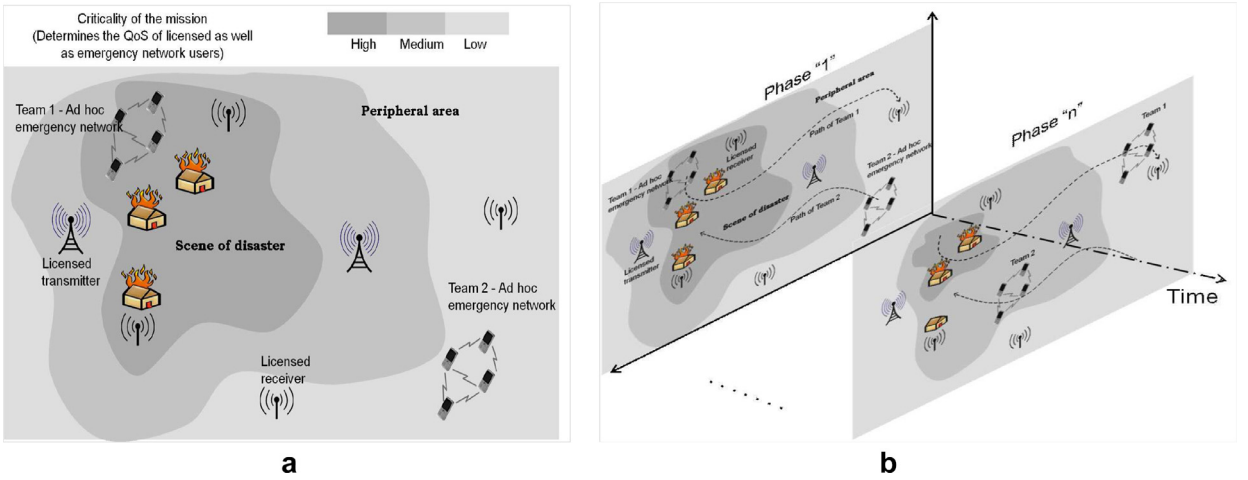


Fig. 1. Emergency networks operating in the vicinity of licensed incumbents in the event of an emergency. Mission policies, which reflect the criticality and hence the QoS of both networks, vary over (a) space (depending on proximity to the scene of the disaster) and (b) over time (depending on the phase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

a discounted price) under the condition that they share it with public safety agencies during emergencies [3,4]. This way, public safety networks will have access to large amount of spectrum resources when required for different types of services (e.g., data messages, real-time voice or video, still picture, and remote control) as well as systems (e.g., self-organization, reliability, scalability, power efficiency, security, multicasting) [5] without the risk of over-subscription while at the same time avoiding undesired under-utilization of the spectrum.

In order for this conditional-auction plan to be viable to licensed users, emergency network operation in licensed spectrum should cause only a *graceful degradation* (to pre-specified levels) in the performance of the incumbents and should not preempt the licensed user traffic at will. In other words, the emergency network should consume only as much spectrum resource as required to achieve a desired level of Quality of Service (QoS) needed for carrying out a mission successfully. In order to achieve this, we envision *a new paradigm for emergency networking*. Emergency networks based on this new paradigm will need to possess the following *cognitive capabilities*: (i) *spectrum agility*, for improving spectrum utilization and robustness against intentional jamming; (ii) *cross layering*, for jointly optimizing communication functionalities; and (iii) *mission-policy awareness*, for steering the emergency network behavior based on the QoS requirements of *both* incumbent and emergency networks.

Currently available solutions proposed for Cognitive Radio (CR) networking [6] in licensed spectrum cannot support our proposed paradigm for emergency networking as they strictly assign priority to the licensed incumbent network over the incoming CR network [7,8]. On the contrary, we envision that the QoS requirements of Emergency Users' (EUs') traffic and that of the Licensed Users' (LUs') of the incumbent network, specified by mission policies, may vary *over space* (from the scene of disaster to the peripheral areas) as shown in Fig. 1(a) and *over time* (during different *phases* of the mission from setup to rescue, recovery, and exit) as shown in Fig. 1(b). The variation in requirements over time can also be attributed to mobility as the ad hoc emergency network traverses through different geographical regions of varying criticality.

An emergency network operating in licensed spectrum while adhering to space- and time-varying mission policies resembles a *multi-agent system*. The multiple autonomous EU agents (the network nodes) of this system try to learn over time the "best behavior", i.e., the choice of transmission parameters that satisfies the QoS of both the incumbent and emergency networks as specified in the high-level mission policies. The controllable transmission parameters that have to be chosen jointly in a cross-layer manner may include signal transmission power, modulation scheme, Forward Error Correction (FEC) type and strength, and Medium Access Control (MAC) parameters.

To overcome the aforementioned problems, firstly, we propose a model-free *Multi-Agent Reinforcement Learning* (MARL)-based [9] communication framework, "RescueNet," for *self-adaptation of coordinating autonomous agents*. Our distributed solution *converges* to the local *optimal joint control policy among EUs* (i.e., optimal choice of transmission parameters at all agents in the neighborhood) through coordination. The optimal control policy ensures *conformance* to the QoS requirements of the emergency and incumbent networks as specified by the high-level mission policy. Secondly, we 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) by adapting the learning parameters on the fly. Thirdly, we propose two novel mechanisms, *bootstrapping* and *selective exploration*, which enable the "experienced" agents to share knowledge with "young" agents in the emergency network in order to expedite the learning process.

This paper is an extended and revised version of one of our prior conference papers [10]. Presently, to the best of our knowledge, there are no *MARL-based emergency networking* solutions. Ours is also the first RL-based

networking solution to exploit the idea of "knowledge sharing" among agents of different ages in order to expedite the learning process. We compare the performance of RescueNet with other popular approaches for solving MARL problems and with a localized optimization approach through extensive simulations [11] in ns-3, a discrete event packet-based simulator [12]. The rest of this paper is organized as follows. In Section 2, we present a summary of prior work on RL-based networking solutions. In Section 3, we propose our solution, the RescueNet framework, which consists of a RL engine that learns and converges over time to a stationary control policy; specifically, in Section 3.1, we provide the necessary background on RL and motivate the need for our framework, while in Section 3.2 we present RescueNet. In Section 4, we evaluate the performance of RescueNet in terms of convergence and conformance. Finally, in Section 5, we draw our conclusions and provide a brief note on future work.

2. Related work

The controllable transmission parameters of the emergency networking nodes have to be chosen jointly in a cross-layer manner so that the QoS requirements of both the incumbent and emergency networks as specified in the high-level mission policies are met. The transmission parameters may include signal transmission power, modulation scheme, FEC type and strength, and MAC as well as routing parameters. An "optimal" choice of parameters may be obtained by solving a centralized cross-layer networking optimization problem based on unrealistic assumptions such as instantaneous knowledge of global network state, complete knowledge of incumbent user performance, and availability of infinite computational capabilities [13–15]. These assumptions compromise optimality apart from rendering the centralized approach impractical.

Another approach to the cross-layer networking is solving a number of localized optimization problems based only on locally observed and shared information [16]. However, this approach has to balance the opposing requirements of capturing local interference constraints as well as satisfying end-to-end (e2e) QoS requirements of the emergency and incumbent network traffic. Traditional networking performance metrics do not capture the interplay among communication functionalities of coexisting networks. Therefore, novel cross-layer performance metrics have to be developed for use in the localized optimization problems [17]. Even though such cross-layer performance metrics may use local observations and models to project and predict e2e behavior and to take local decisions, they cannot guarantee any optimality due to the inadequacy of the prediction models to capture the global network dynamics. Temporary spectrum leasing from licensed users has been studied before in works like [18] under the context of operating secondary mesh networks in primary user spectrum. However, this paradigm, which involves leasing back a portion of the spectrum from the primary user, is only suitable for non-critical, non-emergency commercial operations. Conditional auctioning [4] is better suited for critical public safety operations as the not-for-profit agencies should not be leasing spectrum.

There are some modeling and protocol-based approaches concerned with the problem of efficient and intelligent message forwarding in wireless networks when infrastructure-based communication systems have been damaged or completely destroyed during and in the aftermath of a disaster [19,20]. There are also efforts in applying distributed multi-agent reinforcement learning for wireless networking [21] as well as specific studies on RL for spectrum sensing scheduling and selection in CR mesh networks [22–24], QoS support in Wireless Sensor Networks (WSNs) with and without relay selection [25,26], and sensing coverage [27] in WSNs. However, presently, to the best of our knowledge, there are no *MARL-based emergency networking* solutions. Ours is also the first hybrid distributed RL-based solution that is capable of satisfying both the incumbent as well as the emergency network QoS requirements, and the first to exploit the idea of "knowledge sharing" among agents in order to expedite the learning process. Our solution may look similar to Network Function Virtualization (NFV) [28], Software Defined Network (SDN), or Device to Device (D2D) [29] communications as it employs direct device-to-device communication. However, out solution is a learning-based approach, which adapts the communication parameters on the fly. Our learning-based approach can be one of the solutions for emergency communication as the emergency management faces increasing complexity and decreasing predictability in its operating environment.

3. Proposed solution

RescueNet consists of a RL engine that learns and converges over time to a control policy. Firstly, we provide the necessary background on RL that motivated our choice of a *hybrid* learning approach (distributed yet localized), and then we present our policy-aware emergency networking framework (RescueNet).

3.1. Background and motivation

The underlying concept of RL is finite Markov Decision Process (MDP), which is defined by a tuple $\langle \mathcal{S}, \mathcal{A}, \phi, \rho \rangle$, where $\mathcal{S}$ is a finite set of environment states, $\mathcal{A}$ is a finite set of agent actions, $\phi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ is the state transition probability function, and $\rho : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$ is the reward function. The MDP models an agent acting in an environment where it learns (through prior experiences and short-term rewards) the best control policy π^* (a mapping of states to actions) that maximizes the expected discounted long-term reward. This mapping can be stochastic $\pi : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$ or deterministic $\overline{\pi} : \mathcal{S} \times \mathcal{A} \rightarrow 0 \,||\, 1$.

For deterministic state transition models, the transition probability function ϕ reduces to $\overline{\phi} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow 0 \,||\, 1$ and, as a result, the reward is completely determined by the current state and the action, i.e., $\rho : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$. The state-action pair's goodness value is called *"Q-value,"* and the function that determines the Q-value is called *"Q-function."* An agent can find an optimal control policy

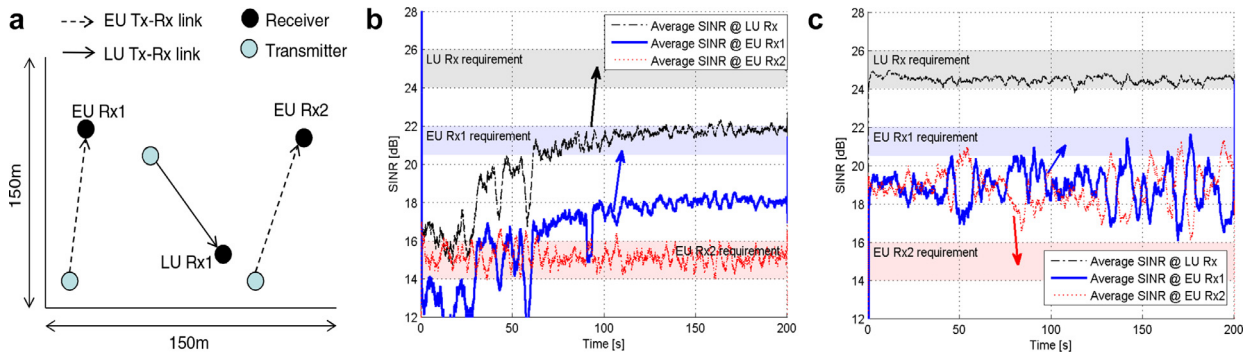


Fig. 2. (a) Topology showing 2 EU transmitter-receiver pairs operating in the vicinity of a LU transmitter-receiver pair; (b) Average SINR at EU and LU receivers when EU transmitters perform transmission power control using *Independent MARL* (only EU Rx2's SINR requirement is met); (c) Average SINR at EU and LU receivers when EU transmitters perform transmission power control using *Global reward MARL* (only LU Rx's SINR requirement is met).

by approximating iteratively its Q-values using prior estimates, short-term reward $r = \rho(s, a) \in \mathbb{R}$, and discounted future reward. This model-free successive approximation technique is called Q-learning. One way to satisfy this criterion is adopting an ϵ-greedy approach where a random action is performed with probability ϵ (*exploration*) and the current knowledge is *exploited* with probability $1 - \epsilon$.

As mentioned earlier, emergency 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 single-agent RL to the multi-agent case is the 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 ϕ and reward function ρ depend on the *joint* environment state and action information, which is not available at any individual agent. Hence, coordination among the autonomous agents is required to achieve fast convergence in a multi-agent scenario. There are three possible approaches to solving MARL problems. We explain each of those approaches with a toy example, the topology of which is depicted in Fig. 2(a), and motivate the need for our hybrid approach, which is then explained in Section 3.2.

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(b) and (c) (20–22 dB for EU Rx1 and 14–16 dB for EU Rx2). The prescribed SINR interval for the LU receiver is also depicted (24–26 dB). *These SINR requirements at both the emergency and the incumbent network nodes represent a simple mission policy specification serving as an example*. The different SINR requirements are derived directly from the corresponding throughput requirements as SINR dictates the achievable channel efficiency in bps/Hz. All the devices operate in the same frequency band (6MHz wide starting at 515 MHz) and the EU transmitters choose from one of the possible five power levels (4–20dBm in steps of 4dB). The transmission power of the LU is fixed at 20dBm. Log-distance path loss model is used to calculate the transmission loss.

Independent MARL [30]: Each agent acts independently *without* coordination. The Q-learning procedure at a node i can be summarized as,

$$Q^i_{n+1}(s, a) = (1 - \alpha^i_n) Q^i_n(s, a) + \alpha^i_n [r + \gamma^i \max_{a' \in \mathcal{A}} Q^i_n(s', a')], \tag{1}$$

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; for simplicity, we consider here a binary reward function (1 if conformed and 0 otherwise). 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. Inability to incorporate information from LUs prevents it from supporting two-sided mission policies, which specify the QoS requirements of both EUs and LUs. Fig. 2(b) shows the SINR at the EU and LU receivers when the EUs try to satisfy their own QoS without any coordination and without the LUs' performance.

Global reward MARL [31]: 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(c).

Distributed Value Function (DVF) MARL [32]: 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. The value iteration procedure at agent i for the state-action pair (s^i, a^i) can be summarized as,

$$Q^i_{n+1}(s^i, a^i) = (1-\alpha^i_n)Q^i_n(s^i, a^i) + \alpha^i_n\left[r^i(s^i, a^i) + \gamma^i \sum_{j\in\mathcal{N}^i} w(i,j)\cdot \max_{a^j\in\mathcal{A}^j} Q^j_n(s^j, a^j)\right], \quad (2)$$

where $w(i, j)$ is the weight that agent i associates with the Q-value estimate obtained from neighboring agent j in the computation of its own Q-value estimate, and $\mathcal{N}^i$ refers to the set of neighboring agents of i. The simplest strategy for computing the weights $w(i, j)$ is to just consider the total number of agents in the neighborhood, i.e., $w(i, j) = 1/|\mathcal{N}^i|$, in which case $\sum_j w(i, j) = 1$. 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 in the neighborhood when incorporated into the value iteration procedure at each agent 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 its effect on the LUs.

Our hybrid learning approach: In order to support effectively 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 in DVF MARL) *as well as* the information of LUs (as in Global reward MARL) obtained from spectrum sensing. The performance of such an approach is shown in Fig. 3. However, the convergence of the hybrid approach exhibits great 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 state-action-pair values exchange). While all of these techniques focus on the conformance, none of them care about convergence to a joint non-detrimental control policy. Our model addresses both the conformance and convergence (through *knowledge sharing*).

3.2. The RescueNet framework

We describe here our specific contributions that will bestow the desired convergence and conformance properties on the RescueNet framework for mobile emergency networking in licensed spectrum. The following are our contributions:

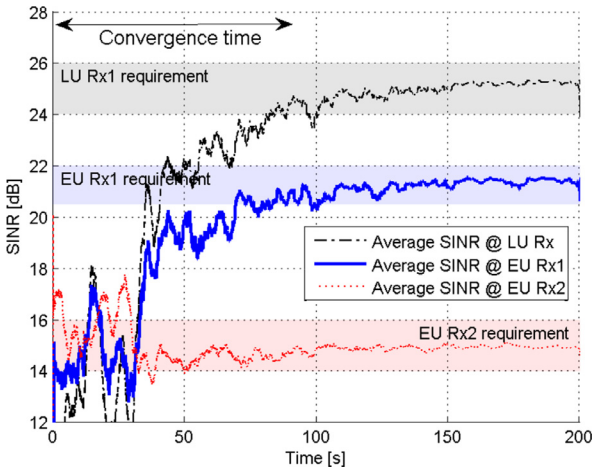


Fig. 3. Average SINR at EU and LU receivers when EU transmitters perform transmission power control using our *hybrid* or *DVF-MARL* approach (LU Rx's SINR requirement and EU Rxs' requirements are met).

- In Section 3.2.1, we cast the 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 both the EU and LU QoS requirements specified by the high-level dynamic mission policies. This forms the core RL engine of RescueNet.
- In Section 3.2.2, we address the significant challenge to the convergence of the learning process posed by the non stationarity of the problem of emergency networking in licensed spectrum. We present mechanisms to adapt the values of key parameters in the iterative approximation of the Q-function to achieve convergence in short time scales.
- In Section 3.2.3, we introduce the novel idea of transferring knowledge from experienced to young agents in the ad hoc network in order to expedite the convergence of young agents to an optimal control policy. We introduce two mechanisms, *bootstrapping* and *selective exploration*, which help expedite the learning process under two different respective scenarios.

3.2.1. Policy-aware emergency networking as a MARL problem

To cast the emergency networking problem as a MARL problem, we identify an individual agent i's states $(\mathcal{S}^i)$, available actions $(\mathcal{A}^i)$, state transition function (ϕ), and reward function (ρ).

States: We represent the state of each 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 MAC and network layers, and k is the destination node to which node i is currently sending data packets. M^i may correspond to a specific time slot, random access delay, or

spreading factor depending on the type of MAC used. Our state transition function is deterministic, i.e., the choice of a certain set of transmission parameters (action) results in a deterministic transition to another state.

Stochastic transitions: In a real-world scenario, however, the transitions will be stochastic in nature. The stochasticity of the system arises due to interference from other concurrent users (based on transmitted power [33]), mobility of users, and non-deterministic channel gains. Each user is also not aware of the number of other users in the system and of their actions. As a result, a user cannot optimize its state (transmission parameters) depending on other users' parameters. We explore alternate MARL formulations where a state is characterized by the average throughput, network delay and jitter, and the user action comprises of change in transmit power, frequency channel, etc. In such cases, a particular action at a user state does not result in a deterministic transition to another state, and the state transition is in fact stochastic. The reward function captures the "goodness" of such transition. Let $T(s, a, s')$ be the probability of transition to state s' from the current state-action pair (s, a). For each (s, a) pair, the reward $r(s, a, s')$ is defined. Let $Q^*(s, a)$ be the expected return for taking action a at a state s and continuing thereafter with the optimal policy, which can be recursively defined as $Q^*(s,a) = \sum_{s' \in S} T(s,a,s')[r(s,a,s') + \gamma \max_{a' \in A} Q^*(s',a')]$. Given the Q values, there is a policy defined by taking, in any situation s, the action a that maximizes $Q(s, a)$. Under the assumption that every situation-action pair is tried infinitely often on an infinite run, the Q values will converge to the true Q^* values [34]. However, for real-world applications like ours, the *exploration vs. exploitation tradeoff* can be leveraged so to converge to the optimal Q^* values without exhaustively searching through the entire state space.

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 EU agent i (source) by incorporating feedback from agent k (destination) about e2e delay (d^{ik}), goodput (gp^{ik}), and SINR of the incumbent network performance (lu). We use goodput instead of throughput as it captures the reliability of data transmission as well. Also, our reward function is generic as any metric (e.g., packet delivery ratio, packet delay, throughput, SINR at the receiver, etc.) that conveys the performance of the incumbent network could also be incorporated without any need for modifications.

$$r^{i,tot} = r^{i,del} + r^{i,gp} + r^{i,lu}; \tag{3}$$

$$r^{i,del} = \begin{cases} 1 - \dfrac{d^{ik} - d^{min}}{d^{max} - d^{min}}, & d^{min} \le d^{ik} \le d^{max} \\ a, & d^{ik} < d^{min} \\ b, & d^{ik} > d^{max} \end{cases}; \tag{4}$$

$$r^{i,gp} = \begin{cases} 1 - \dfrac{gp^{max} - gp^{ik}}{gp^{max} - gp^{min}}, & gp^{min} \le gp^{ik} \le gp^{max} \\ a, & gp^{ik} > gp^{max} \\ b, & gp^{ik} < gp^{min} \end{cases}; \tag{5}$$

$$r_n^{i,lu} = \begin{cases} 1 - \dfrac{lu - lu^{min}}{lu^{max} - lu^{min}}, & lu^{min} \le lu \le lu^{max} \\ a, & lu < lu^{min} \\ b, & l > lu^{max} \end{cases}. \tag{6}$$

Eqs. (4) and (5) show how the reward function captures the requirements for EUs, and (6) shows how the reward function captures the requirements for LUs as specified by the mission policies. The positive reward for delay performance is high (i.e., close to the maximum reward value of 1) if 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 with the mission policy's requirements. The magnitude of the rewards (in conjunction with the learning and discount factors) are chosen in such a way to ensure that the Q-value estimates do not vary too much within ± 2 db with a single reward. In (4)–(6), $-1 < a, b < 0$ and $a > b$.

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 met. Note that negative rewards are given when the experienced goodput exceeds the maximum threshold value and when the experienced delay is below the minimum threshold value [35]. The philosophy behind these negative rewards for exceeding the requirements is that the emergency network should consume only as much spectrum resource as required to achieve a desired level of QoS needed for carrying out a mission successfully. Exceeding the requirements penalizes the other nodes in the emergency as well as in the incumbent networks.

Incorporation of mission policy: The goodput and delay thresholds gp^{min}, gp^{max}, d^{min}, and d^{max} together give RescueNet the flexibility to support four different traffic classes, namely, (i) *loss tolerant and delay tolerant* (e.g., scalar data from sensors and multimedia content such as snapshots, which are non time critical), (ii) *loss tolerant and delay sensitive* (e.g., video streams that can be within a certain level of distortion), (iii) *loss sensitive and delay tolerant* (e.g., critical data that requires offline post processing), and (iv) *loss sensitive and delay sensitive* (e.g., time-critical mission directives and alerts).

Inverse reinforcement learning (future work): The scalar reward function does not provide optimal performance in dynamically changing environment as the reward function is static. Hence, we will study and formulate the inverse reinforcement learning problem to optimize the

reward function when apriori knowledge is available on the fly. The inverse RL problem consists in finding a reward function that can explain observed behavior [35]. Thus, if we know the measurements of an agent's behavior (i.e., goodput and delay) over time in a variety of circumstances (i.e., number of users, required bandwidth, and available channels), we formulate and optimize the reward function based on given circumstances. We will focus initially on the setting in which the complete prior knowledge and mission policy are given; then, we will find new methods to choose among optimal reward functions as multiple possible functions may exist. We will also study the improvement in performance (in terms of goodput and delay) when flexible reward (with inverse RL) functions are used to incorporate robustness in the learning process and thus handling fluctuations.

3.2.2. Convergence under non stationarity

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 to the convergence of our MARL problem posed by this non-stationarity by considering segments of the MARL problem over time as a repeated static game [9], where the centralized optimization of transmission parameters is not feasible. In a static game, the rewards depend only on the joint choice of the transmission parameters of the nodes and, hence, the control policy transforms into $\pi : \mathcal{A} \rightarrow [0, 1]$. The game is referred to as a repeated static game as it is played repeatedly over time by the same set of nodes. However, stabilization of the learning procedure in this repeated static game requires a balancing of the exploration–exploitation trade-off and an appropriate choice of learning factor. Note that we have used the metric, *conformance rate*, the percentage of time spent in conformance with the mission policy, to measure indirectly the convergence time because there is no guarantee on convergence in such a non-stationary environment. Incorrectly-chosen learning factors may tamper the sensitivity of the proposed solution; for this reason, we have carefully chosen those values to accommodate environmental changes (degree of mobility) based on our empirical study upon simulations. As it is hard for the reinforcement learning framework to manage multiple parameters, we have introduced "bootstrapping" and "selective exploration", which are detailed below, in order to reduce the sensitivity of the proposed solution.

Exploration-exploitation trade-off: In RescueNet, the exploration factor ϵ (of the ϵ-greedy approach) is time varying with a high value in the beginning of each static game (*more exploration*) and a with low value at the end of each static game (*more exploitation*). When changing the transmission parameters, RescueNet selects random (but selective) parameters for exploration, and selects the optimal parameters for exploitation. The exploration factor ϵ is a normalized number ranging from 0 to 1. We determine the exploration decay rate δ_ϵ of the exploitation factor at all agents based on the degree of mobility, i.e., $\delta_\epsilon = \psi(v)$, where v is the average speed of all the nodes in the emergency network. An estimate of the average speed of nodes can be obtained from the nature of the mission the team of nodes is involved in (first response, rescue, recovery, or exit). In case of low mobility, nodes should exploit their knowledge more as their environment changes very slowly. In the case of medium node mobility, nodes should explore more than they exploit as their acquired knowledge may become outdated sooner than in the case of low node mobility. However, in case of very high mobility the environment may change sooner than the time the RL engine takes to converge. The evolution of the exploration factor over time is given by $\epsilon^i_{n+1} = \epsilon^i_n \cdot \delta_\epsilon$. However, once the exploration factor reaches a low value, it is reset to the initial value in order to ensure that the learning process does not cease.

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, as shown in (2). 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 the 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 high node mobility, nodes should refrain from using their experience as it may be outdated. In case of low mobility, nodes should exploit their knowledge more as their environment changes very slowly. The time evolution of learning factor is given by $\alpha^i_{n+1} = \alpha^i_n \cdot \delta_\alpha$. Similar to the exploration factor, in order to ensure that the learning process does not cease once the learning factor reaches a low value, it is reset to its initial value.

3.2.3. Knowledge sharing among agents

RescueNet, with its time-varying learning parameters (α and ϵ) can enable convergence of multiple agents to an optimal joint control policy. However, the convergence takes time as the process of Q-learning requires exploration of all possible 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 two novel mechanisms for knowledge sharing among agents, *bootstrapping* and *selective exploration*. To understand these concepts better, consider the following examples.

Bootstrapping: Team 1 in Fig. 1 is working in the scene of disaster, while Team 2 is moving from the peripheral area towards Team 1 to replace it. The agents of Team 1 are "experienced" as far as the "scene of disaster" is concerned, while the agents of Team 2 are "young." Once the agents of Team 2 reach the new region, each of them broadcasts a request for knowledge (i.e., Q-value and its state-action-pair values) from the experienced agents. The agents of Team 1 that receive this request start a countdown timer, the duration of which depends on the degree of proximity (a normalized metric computed based on pre-specified minimum and maximum distances) to the requesting agent. Once the first response is sent out by the closest agent in Team 1, the other experienced agents abort

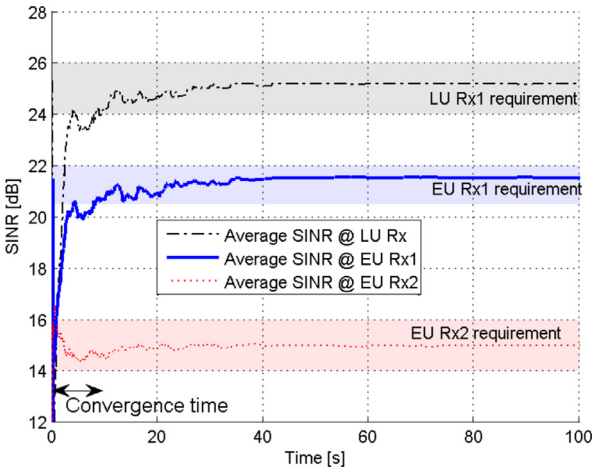


Fig. 4. Average SINR at EU and LU receivers when EU transmitters perform transmission power control using our hybrid approach along with bootstrapping, i.e., knowledge of good initial states for the learning process (both LU and EU Rx's SINR requirements are met).

their timers and mark that request as expired. This bootstrapping allows the agents of Team 2 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 (shown in Fig. 4) than they would have under usual circumstances (shown in Fig. 3).

Selective exploration: Consider another example where Team 2 is moving from the peripheral area towards Team 1 to form a bigger team with more data traffic. Once the agents of Team 2 reach the new region and broadcast requests for knowledge from experienced agents, the agents of Team 1, who are already aware of the levels of intra- and inter-network interference, provide guidelines for selective exploration strategies to the new learning agents. These selective exploration guidelines prevent the new agents from exploring already infeasible states (such as the ones corresponding to high power levels in frequency bands used by LUs and many EUs), where $T(s, a, s') = 0$. Selective exploration again reduces the time to converge to an optimal joint control policy. Note that bootstrapping and selective exploration both fall under our knowledge-sharing mechanism, which is aimed at expediting the learning process. Selective exploration using stochastic transition and overall RescueNet algorithm for a node i is summarized in Algorithm 1.

Algorithm 1 RescueNet for Agent i.

Initialize:
$t = 0, \alpha, \gamma$, and ε
initialize starting state s_t^i
if Bootstrapping then
  Copy the Q-value $Q(s, a, s')$ of the replacement node
else
  for each $s \in \boldsymbol{S}$, $a \in \boldsymbol{A}$, $s' \in \boldsymbol{S'}$ **do**
    initialize the Q-value $Q(s, a, s')$
  end for
end if
Learning:
while do
  generate random number *rand* between 0 and 1
  if $(rand < \varepsilon)$ **then**
    Selective Exploration: select the action a_t^i based on the transition probability T
  else
    Exploitation: select the action a_t^i characterized by the maximum Q-value
  end if
  waiting for feedback
  calculate aggregated reward($r_{t\alpha}^i$) based on the policy and feedback
  update Q-value using Eq. 2
  $s_t^i = s_{t+1}^i$
  (decay α and ε for convergence within a static game)
end while

4. Performance evaluation

In order to evaluate the performance of RescueNet, we implemented it on ns-3, a packet-based discrete-event network simulator [12], and performed three different campaigns of simulations. In this section, we explain the objective of each campaign, report the individual simulation's settings and assumptions (in terms of data traffic and mission policy), and provide our observations.

In *Campaign I*, we performed simulations to study the performance of RescueNet in a controlled static setting. This campaign is different from what is presented in Section 3.1 as this deals with a much larger state space (tunable transmission parameters) and the mission policies are represented using entirely different e2e metrics (goodput and delay). Specifically, the simulations are aimed (i) at comparing the performance of RescueNet with other frameworks in terms of conformance to a specified high-level mission policy and (ii) at demonstrating how RescueNet adapts to changes in mission policy over time.

In *Campaign II*, we performed simulations with node mobility to show that the performance of RescueNet is not dependent on any specific network topology and that it adapts to the non-stationarity in the environment. Specifically, the simulations are aimed (i) at showing RescueNet's ability to adapt to changes in mission policies over time *and* space under mobility and (ii) at discussing the importance of adapting the key learning parameters in RescueNet for convergence and conformance in a non-stationary environment.

In *Campaign III*, we performed simulations to demonstrate the effectiveness of knowledge sharing. Specifically, the simulations are aimed (i) at studying the benefits of knowledge transfer among agents (Q-tables in order to help the new agents *bootstrap*) when a new flow replaces an existing flow in terms of speed of convergence to an optimal joint control policy and (ii) at studying the merits of *selective exploration* when a new flow is added to the existing data traffic in an EU team.

Transmission parameters and assumptions: The tunable transmission parameters and assumptions regarding the loss model and the MAC scheme considered are listed in Table 1. In all the simulations the EU nodes employ a Direct Sequence Code Division Multiple Access (DS-CDMA) MAC with self-assigned variable-length chaotic codes [36] and the Most Forward within Radius (MFR) geographical routing scheme [37]. Log-distance path loss model is used to calculate the transmission loss. We assumed there is

Table 1
Simulation settings (tunable transmission parameters) and assumptions.

Transmission power	4–20 dBm in steps of 4 dB
Transmission band	3 channels in 515–533 MHz band (each 6 MHz wide)
Modulation scheme	8-, 16-, 32-QAM
MAC	DS-CDMA with chaotic spreading codes
Loss model	Log-distance path loss model

a shared network to exchange light-weight control messages such as neighborhood discovery protocol and resource management protocol and that their overhead is in the order of 100 bytes/s to run RescueNet.

Conformance with the QoS requirements of the LUs of incumbent network requires either explicit feedback from the LUs themselves (full observability) or requires knowledge of estimates of the worst-case incumbent network performance at certain emergency nodes based on their distance from incumbent transmitters (partial observability). To acquire information about the incumbent users, we assumed that the emergency-network nodes employ cyclostationary feature detection [38,39]. This capability ensures that the EUs cannot only detect the presence of data traffic in certain frequency bands but also differentiate between LU and EU traffic.

4.1. Campaign I

The topology of EU and LU nodes used in this campaign of simulations is depicted in Fig. 5, which shows two teams of EUs operating in the vicinity of a LU receiver.

(1) Conformance to the mission policy. We compared RescueNet with (i) a framework that employs the localized optimization approach similar to the ones proposed in [16,17] (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 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.

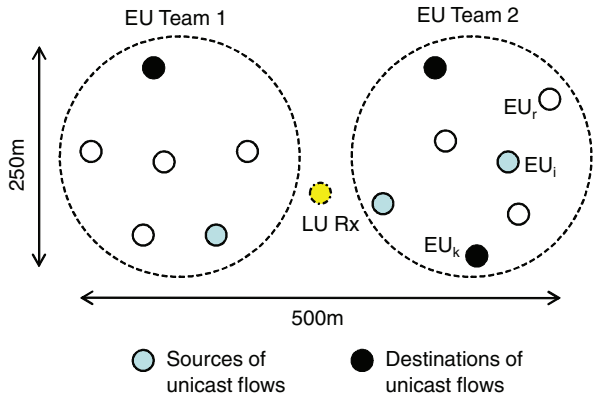


Fig. 5. Scenario of two EU teams operating in the vicinity of a LU Rx used to evaluate RescueNet in terms of conformance to mission policy and convergence to an optimal control policy.

The mission policy: Both teams of EUs try to comply with the QoS requirements of the same mission policy. The QoS requirement of LUs is specified in terms of acceptable average received SINR values as it can be estimated at any EU with spectrum sensing capabilities. The QoS requirement of the emergency network is specified in terms of acceptable application layer goodput and packet delay. 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. The SINR requirement at the LUs was set to $lu^{min} = 25, lu^{max} = 33$dB, and the goodput and delay requirements of the EUs were set to $gp^{min} = 480, gp^{max} =$ 510Kbps and $d^{min} = 6, d^{max} = 12$ms, respectively. The policy requirements are shown as blue-shaded regions in Fig. 6. The values of a and b in the reward functions (4)–(5) were set to $a = -0.25$ and $b = -0.5$. The results presented in Fig. 6 are based on 50 independent trials performed to achieve a relative confidence interval $< 5\%$ so to give statistical relevance to the results.

Observations: Fig. 6(a) and (b) shows the aggregated goodput and average packet delay, respectively, of all the three unicast flows in the emergency network. Fig. 6(c) shows the average SINR measured at the LU. It can be observed that the emergency network fully conforms to

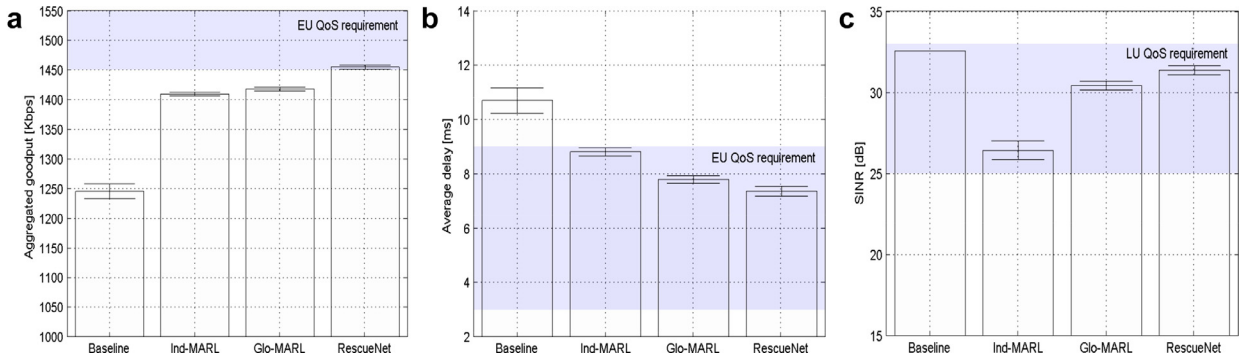


Fig. 6. Campaign I: Emergency network performance in terms of (a) aggregated goodput, (b) average packet delay, and (c) average SINR at LU receivers when EU nodes employ a local optimization approach, independent MARL (Ind-MARL), global reward MARL (Glo-MARL), and RescueNet. (For interpretation of the references to color in this figure in text, the reader is referred to the web version of this article.)

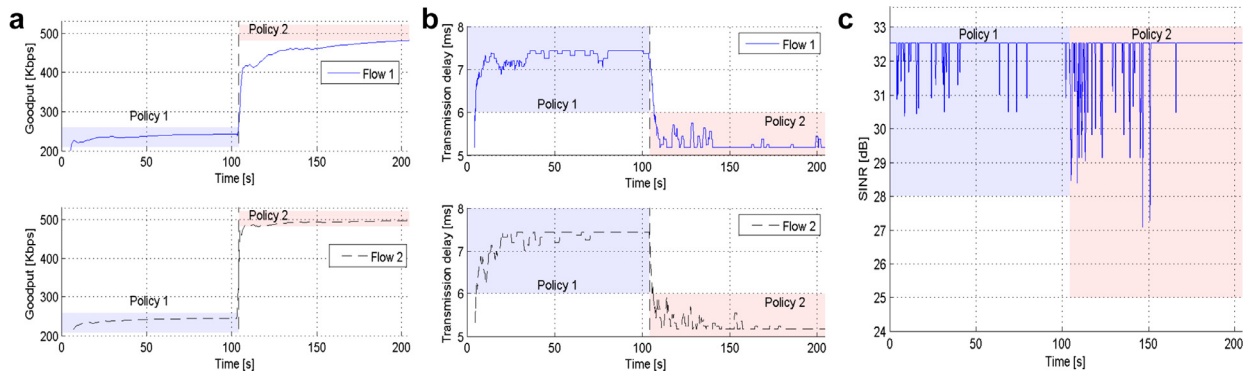


Fig. 7. Campaign I. RescueNet's ability to adapt to time-varying mission policies (a) Goodput of flows 1 and 2; (b) Average packet delay of flows 1 and 2; and (c) Average SINR at the LU receiver when the mission policy specification changes over time. (For interpretation of the references to color in this figure in text, the reader is referred to the web version of this article.)

the mission policy specification 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 on the incumbent nodes. As a result, the EU unicast flows suffer from huge 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. 6(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 based 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, besides not guaranteeing any optimality, Baseline also incurs a huge overhead. A node using a baseline approach requires at least 8 bytes of data, i.e., transmission power (2), channel (2), modulation (2), and location (4) from its neighboring nodes, whereas RescueNet requires only 4 bytes of data, i.e., maximum reward from the LU (2) and EU (2). Independent MARL does not incur any communication overhead, Global MARL incurs 2 bytes for maximum reward for LU, and DVF-MARL incurs 2 bytes overhead for maximum reward for EUs.

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 with their own QoS requirements but also take into account the effect of their actions on their neighbors. This exchange incurs only a negligible overhead as this small amount of information can be piggy-backed with the frequent control packets that any other functionality may already require (e.g., neighborhood discovery). RescueNet also enables conformance with the requirements of the incumbent receivers as it incorporates estimates of the worst-case LU performance and incorporates the same in the reward function.

(2) Adaptation to time-varying mission policy. One of the main attributes of RescueNet is its ability to conform to time-varying mission policies. To verify this ability, we use the setup depicted in Fig. 5 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. The EUs and LUs operate in the same frequency band, i.e., there is only one channel for use.

The time-varying mission policy: EUs try to satisfy the QoS requirements specified by Policy 1 at the beginning of the experiment. The data traffic in the EU team at this time was set to two unicast flows, 250 Kbps each. Initially, the SINR requirement at the LUs was set to $lu_1^{min} = 28, lu_1^{max} = 33$dB and the goodput and delay requirements of the EUs were set to $gp_1^{min} = 230, gp_1^{max} = 260$Kbps and $d_1^{min} = 6, d_1^{max} = 8$ms, respectively. After 100 s into the experiment, the increase in priority of emergency network over the incumbent network was simulated by increasing the rate of both the unicast flows to 500 Kbps. Policy 2 was enforced by changing the goodput and average delay requirements of the EUs to $gp_2^{min} = 480, gp_2^{max} = 510$Kbps and $d_2^{min} = 5, d_2^{max} = 6$ms, respectively. The SINR requirement of the LU was reduced to $lu_2^{min} = 25, lu_2^{max} = 33$dB according to this new policy. The two different policy requirements are shown as pink- and gray-shaded regions in Fig. 7.

Observations: Fig. 7(a) and (b) shows the average (moving window) goodput and packet delay, respectively, of each unicast flow in the team of EUs. Fig. 7(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 imposed by the two different mission policies. This flexibility is due to the generic nature of the reward function of the proposed RescueNet framework. The maximum and

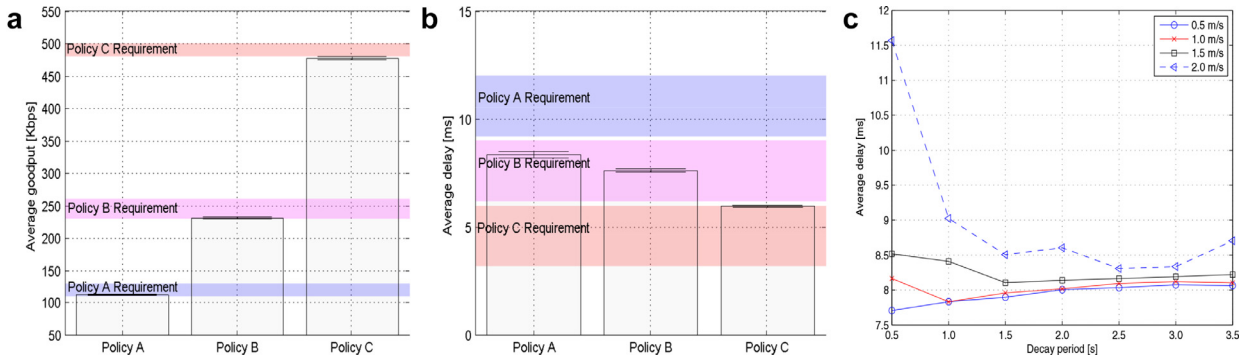


Fig. 8. Campaign II. (a) Average goodput and (b) delay of flows corresponding to policies A, B, and C when 3 mobile teams under 3 distinct policies are operating in the vicinity of a LU pair. Average node speed is 1 m/s; (c) Impact of the periodicity of decay of learning factor and of growth of exploitation factor on conformance with a mission policy under different node velocities.

minimum limits in all the three components of the reward function (4)–(6) can be varied over time to capture the QoS requirements of different mission policies.

4.2. Campaign II

To obtain results that help demonstrate conclusively RescueNet's ability to adapt to the non stationarity in the operating environment, we performed simulations with node mobility as well as with time- and space-varying mission policies. These simulations were also intended to show that RescueNet's performance is not dependent on any specific topology of emergency network nodes.

*Mobility pattern:*The EUs in all teams perform a random walk (for a randomly chosen duration between 5 and 20 s) within a rectangular area (200×200 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 and random-pause durations. This was done to eliminate any bias that may be introduced by a static network topology.

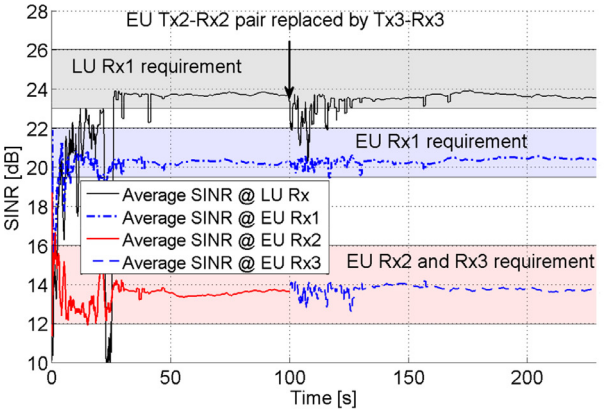
(1) Policy conformance under mobility. We considered a simulation scenario with 3 teams (with five nodes per team and with one unicast flow per team) of EUs operating in the vicinity of a LU pair. At any given point in time, the 3 teams adhere to distinct mission policies. In addition, a team operates under different policies from time to time over the course of the simulation. This was done to eliminate any bias to a specific combination of policies among the 3 teams. Also, the results we obtained were averaged over 50 trials – with the sequence of events differing in each trial – to obtain very small relative confidence intervals. In the simulations, the average speed of nodes was chosen in a uniform random manner between 0.5 and 1.5 m/s. The EUs decide on the 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 RescueNet.

*Time-varying mission policy:*The QoS requirements (goodput and delay) of the three policies A, B, and C were set to $gp_A^{min} = 115, gp_A^{max} = 130$Kbps and $d_A^{min} = 9, d_A^{max} = 12$ms, $gp_B^{min} = 230, gp_B^{max} = 260$Kbps and $d_B^{min} = 6, d_B^{max} = 9$ms, and $gp_C^{min} = 480, gp_C^{max} = 510$Kbps and $d_C^{min} = 3, d_C^{max} = 6$ms, respectively, in increasing order of QoS levels. The corresponding LU SINR requirement for the three polices were $lu_A^{min} = 30, lu_A^{max} = 33$dB, $lu_B^{min} = 28, lu_B^{max} = 33$dB, and $lu_C^{min} = 25, lu_C^{max} = 33$dB. The offered load corresponding to policies A, B, and C were set to 125, 250, and 500 Kbps, respectively.

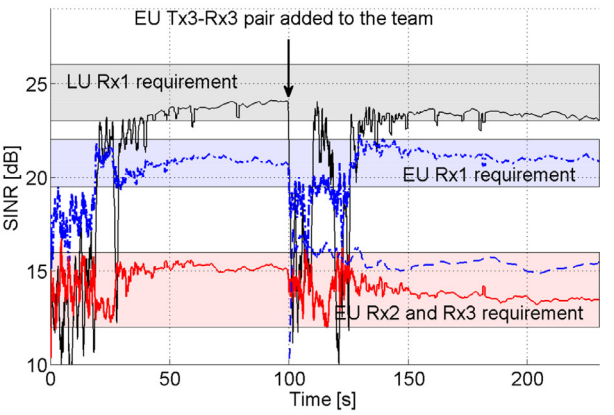
Observations: Our simulation results in Figs. 8(a) and (b) show clearly that the average goodput and average delay of unicast sessions corresponding to the three different mission policies A, B, and C are very close to the goodput and delay specifications of each of those policies with small relative 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 the non stationarity in the environment.

(2) Adaptation of learning parameters. To adapt to the non stationarity of the environment (due to node mobility), RescueNet needs to determine dynamically the values of its learning parameters (α and ϵ) and their decay/growth rates (δ_α and δ_ϵ) based on the degree of node mobility. However, prior knowledge about the relationship between the degree of mobility and the *periodicity of decay/growth of the learning parameters* is essential to adapt the rates online for convergence and conformance.

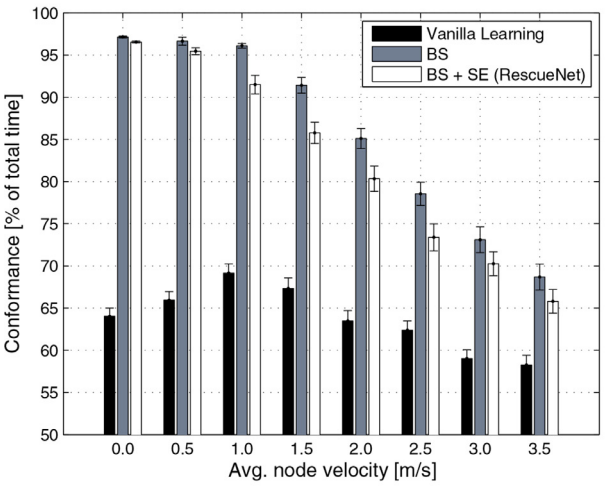
Observations: Fig. 8(c) compares the performance of the emergency network in terms of average packet delay at various decay periodicity values for four different average node velocities (0.5, 1.0, 1.5, 2.0 m/s). We can observe that, as the node velocity increases, the decay period has to be increased to achieve delays that conform with 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. This ability to learn quickly can be retained only if the decay period of α is high. This offline tuning of the periodicity is essential to choose the right values of learning parameters and their



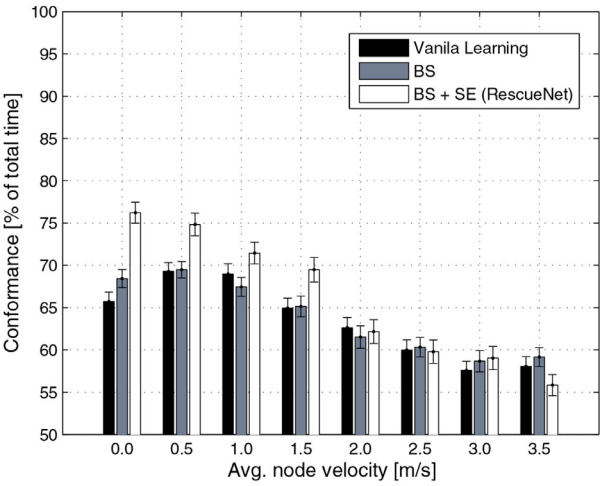
(a) Replacement of an EU Tx-Rx pair (single instance)



(b) Addition of an EU Tx-Rx pair (single instance)



(c) Replacement of an EU Tx-Rx pair



(d) Addition of an EU Tx-Rx pair

Fig. 9. Campaign III. Effect of Bootstrapping (BS) and Selective Exploration (SE) over time when an EU pair in the team is (a) replaced and (b) added, respectively. Conformance (in terms of % of total time, here 100 s) of the EU team in the vicinity of a LU pair (c) when one EU Tx-Rx pair is replaced by a new pair and BS is employed, and (d) when a new Tx-Rx pair is added to the EU team and SE is employed. *Vanilla Learning* refers to the hybrid learning approach without any knowledge sharing.

decay/growth rates for ensuring policy conformance even under mobility.

4.3. Campaign III

RescueNet enables knowledge sharing (bootstrapping and selective exploration depending on the scenario) among agents of different ages in order to reduce the convergence time to the optimal joint control policy. To verify this ability we use the following setup. Team 1 with two EU Tx-Rx pairs is operating in the vicinity of one LU Tx-Rx pair. The EUs decide on the appropriate values of the following transmission parameters in a cross-layer manner: frequency band and transmission power level using our hybrid learning mechanisms.

Mobility pattern: Each EU performs a random walk within a rectangular area ($20 \times 20\text{m}^2$) around their initial positions with a pause after every walk. The velocity is increased up to 3.0 m/s in steps of 0.5 m/s. This mobility pattern simulates movement patterns of first responders in the scene of disaster. This was done to eliminate any bias that may be introduced by a static network topology.

(1) Bootstrapping. In order to demonstrate the effectiveness of bootstrapping, we assume that a new EU Tx-Rx pair belonging to a different team replaces one of the EU Tx-Rx pairs in Team 1 resembling a replacement of overworked first responders involved in a specific mission over time.

Mission policy: EU pairs operating in the vicinity of a LU pair perform transmission power control to ensure that the SINRs at their receivers EU and LU receivers are within prescribed intervals (20–22 dB for EU Rx1 and 14–16 dB for EU Rx2). The different SINR requirements are derived directly from the corresponding throughput requirements as SINR dictates the achievable channel efficiency in

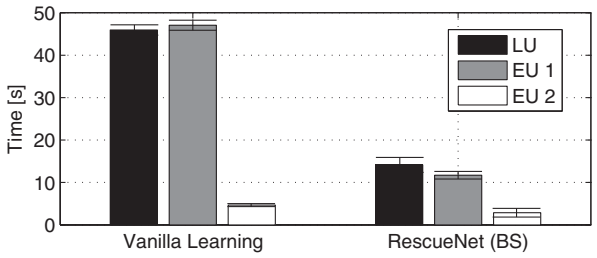


Fig. 10. Convergence time of vanilla learning and RescueNet (with its bootstrapping mechanism). Avg. node velocity is set 0.5 m/s. Convergence time is defined as the time t when the SINR of *all* the channels conforms to their mission policy with a tolerance rate of 5% afterward.

bps/Hz. In the bootstrapping experiment, the new EU pair replaces EU Tx2-Rx2 pair and retains the same policy specification.

Observations: The time taken by all the Tx-Rx pairs in Team 1 to converge back to an optimal joint policy and, hence, the percentage of time spent in conformance with the mission policy, were measured by observing the behavior (from multiple runs for statistical relevance) over time as shown in Fig. 9(a) and (b). It is evident from Fig. 9(c) and (d) that RescueNet with its bootstrapping mechanism conforms to the mission policy 30% more time than the vanilla hybrid learning mechanism. Fig. 10 shows that RescueNet converges nearly 70% faster than the vanilla hybrid learning mechanism. It is because the transfer of additional knowledge (from the experience to the young agents) of the appropriate initial state expedites the learning process.

(2) Selective exploration. In order to demonstrate the effectiveness of selective exploration, we assume that a new EU Tx-Rx pair belonging to a different team joins Team 1 thus increasing the total number of EU Tx-Rx pairs resembling a scenario where reinforcement is called in to increase the team size.

Mission policy: The mission policy (in terms of SINR requirements of the LUs and EUs) for the selective exploration experiment is the same as that of the bootstrapping experiment with a minor change: the new EU pair, which is added to Team 1, has the same SINR requirement as the EU Tx2-Rx2 pair.

Observations: In Fig. 9(d), which corresponds to the scenario where a new EU Tx-Rx pair is added to Team 1, it can be seen that applying bootstrapping solely to the new Tx-Rx pair is ineffective as the interference pattern of the entire network is affected. This change in interference map necessitates learning at all the agents and "selective exploration" at all agents serves to expedite convergence. Selective exploration avoids unnecessary exploration of states already deemed infeasible by the experienced agents. Selective exploration has its limitation under high degree of node mobility in the network (e.g., when the average velocity of nodes is greater than 2.0 m/s).

(3) Scalability. To show scalability and feasibility of RescueNet, we increase the number of EUs in the existing incumbent network.

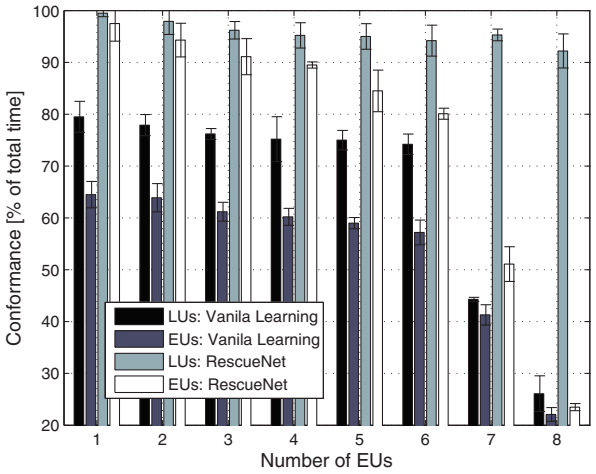


Fig. 11. EU's and LU's performance of Vanilla Learning and RescueNet (BS+SE) when increasing the number of EU pairs in the existing incumbent network with 5 LU pairs.

Mission policy: EU pairs operating in the vicinity of a LU pair perform transmission power control to ensure that the SINRs at their receivers are within prescribed intervals: (20–22 dB, 14–16 dB, and 24–26 dB). The prescribed intervals are randomly selected for EUs and LUs in the simulations. The different SINR requirements are derived directly from the corresponding throughput requirements as SINR dictates the achievable channel efficiency in bps/Hz.

Observations: Fig. 11 shows EU's and LU's performance of Vanilla Learning and RescueNet (BS+SE) when increasing the number of EU pairs in the existing incumbent network with 5 LU pairs. This simulation is designed to assess the maximum number of users and performance degradation of EUs and LUs when the number of EUs increases. It shows that in a confined space ($600 \times 600\text{m}^2$), the performance of EUs drops drastically when more than 6 EUs pairs come into the field, while the performance of LUs is marginally compromised. We can imply that our solution does not degrade the performance of LUs due to the selective exploration even if the number of users exceeds the capacity of channel.

5. Conclusion and future work

We introduced a policy- and learning-based paradigm for emergency networking in conditionally auctioned licensed spectrum. The concept of mission policies, which specify the Quality of Service (QoS) for emergency as well as for incumbent network traffic, is envisioned. Our paradigm for emergency networking represents a shift from the established primary–secondary model (which uses fixed priorities) and enables graceful degradation in the QoS 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. The proposed solution can go beyond the emergency scenario and has the potential to enable cognitive ad hoc network operation in

any frequency band, licensed or unlicensed. The performance of RescueNet in terms of convergence and policy conformance is verified using simulations on ns-3.

Our future work includes Inverse Reinforcement Learning (IRL) for the reward function. As the scalar reward function does not provide optimal performance in dynamically changing environment, we will study the IRL problem to optimize the reward function when a priori knowledge is available on the fly. The IRL problem consists in finding a reward function that can explain observed behavior. We will focus initially on the setting in which the complete prior knowledge and mission policy are given; then, we will find new methods to choose among optimal reward functions as multiple possible reward functions may exist.

References

[1] The First Responder Network (FirstNet) and Next-Generation Communications for Public Safety: Issues for Congress, (https://www.fas.org/sgp/crs/homesec/R42543.pdf).

[2] K. Mase, How to deliver your message from/to a disaster area, Commun. Mag., IEEE 49 (1) (2011) 52–57.

[3] Public-Private Partnership Options for Managing Wireless Networks, (http://research.policyarchive.org/19928_Previous_Version_2007-06-25.pdf).

[4] C. Bazelon, Too many goals: problems with the 700 MHz auction, Inf. Econ. Policy (Special Section on Auctions) 21 (2) (2009) 115–127.

[5] P. Pawelczak, R. Venkatesha Prasad, L. Xia, I. Niemegeers, Cognitive radio emergency networks - requirements and design, in: Proceedings of IEEE Symposium of New Frontiers in Dynamic Spectrum Access Networks (DySPAN), 2005, pp. 601–606.

[6] J. Mitola, G.Q. Maguire, Cognitive radio: making software radios more personal, IEEE Personal Commun. Mag. 6 (4) (1999) 13–18.

[7] IEEE 802 LAN/MAN Standards Committee 802.22 WG on WRANs (Wireless Regional Area Networks), (http://www.ieee802.org/22/).

[8] Y. Yuan, P. Bahl, R. Chandra, P. Chou, J. Ferrell, T. Moscibroda, S. Narlanka, Y. Wu, Knows: cognitive radio networks over white spaces, in: Proceedings IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN), Dublin, Ireland, 2007.

[9] L. Busoniu, R. Babuska, B. De Schutter, A comprehensive survey of multiagent reinforcement learning, IEEE Trans. Syst., Man, Cybern., Part C: Appl. Rev. 38 (2) (2008) 156–172.

[10] E.K. Lee, H. Viswanathan, D. Pompili, Learning-based framework for policy-aware cognitive radio emergency networking, in: Proceedings of IEEE Global Communications Conference (GLOBECOM), 2013, pp. 968–973.

[11] RescueNet module for ns3, (http://db.tt/iHt86OfQ).

[12] ns3: Network Simulator, (http://www.nsnam.org/).

[13] G. Cheng, W. Liu, Y. Li, W. Cheng, Joint on-demand routing and spectrum assignment in cognitive radio networks, in: Proceedings of IEEE International Conference on Communications (ICC), Glasgow, UK, 2007.

[14] Y. Hou, Y. Shi, H. Sherali, Optimal spectrum sharing for multi-hop software defined radio networks, in: Proceedings of IEEE Intl. Conference on Computer Communications (INFOCOM), Anchorage, AK, 2007.

[15] L. Lai, H. El-Gamal, H. Jiang, H. Poor, Cognitive medium access: exploration, exploitation, and competition, IEEE Trans. Mobile Comput. 10 (2) (2011) 239–253.

[16] L. Ding, T. Melodia, S. Batalama, J. Matyjas, M. Medley, Cross-layer routing and dynamic spectrum allocation in cognitive radio ad hoc networks, IEEE Trans. Veh. Technol. 59 (4) (2010) 1969–1979.

[17] D. Pompili, I.F. Akyildiz, A multimedia cross-layer protocol for underwater acoustic sensor networks, IEEE Trans. Wirel. Commun. 9 (9) (2010) 2924–2933.

[18] S. Shakeri, A. Sayegh, T. Todd, Secondary wireless mesh network design using leased frequency spectra, in: Proceedings of Wireless Communications and Networking Conference (WCNC), 2010.

[19] P. Kolios, A. Pitsillides, O. Mokryn, K. Papadaki, Explore and exploit in wireless ad hoc emergency response networks, in: Proceedings of IEEE International Conference on Communications (ICC), 2014, pp. 452–458.

[20] S. George, W. Zhou, H. Chenji, M. Won, Y.O. Lee, A. Pazarloglou, R. Stoleru, P. Barooah, Distressnet: a wireless ad hoc and sensor network architecture for situation management in disaster response, Commun. Mag., IEEE 48 (3) (2010) 128–136.

[21] C.-K. Tham, J.-C. Renaud, Multi-agent systems on sensor networks: a distributed reinforcement learning approach, in: Proceedings of Intelligent Sensors, Sensor Networks and Information Processing Conference (ISSNIP), 2005.

[22] M. Di Felice, K. Chowdhury, A. Kassler, L. Bononi, Adaptive sensing scheduling and spectrum selection in cognitive wireless mesh networks, in: Proceedings of Computer Communications and Networks (ICCCN), 2011.

[23] K. Chowdhury, R. Doost-Mohammady, W. Meleis, M. Di Felice, L. Bononi, Cooperation and communication in cognitive radio networks based on tv spectrum experiments, in: Proceedings of IEEE World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2011, pp. 1–9.

[24] J. Lunden, S. Kulkarni, V. Koivunen, H. Poor, Multiagent reinforcement learning based spectrum sensing policies for cognitive radio networks, Select. Top. Signal Process., IEEE J. 7 (5) (2013) 858–868.

[25] X. Liang, M. Chen, Y. Xiao, I. Balasingham, V. Leung, A novel cooperative communication protocol for QoS provisioning in wireless sensor networks, in: Proceedings of Testbeds and Research Infrastructures for the Development of Networks Communities and Workshops (TridentCom), 2009.

[26] X. Liang, I. Balasingham, V. Leung, Cooperative communications with relay selection for QoS provisioning in wireless sensor networks, in: Proceedings of Global Telecommunications Conference (GLOBECOM), 2009.

[27] J.-C. Renaud, C.-K. Tham, Coordinated sensing coverage in sensor networks using distributed reinforcement learning, in: Proceedings of IEEE International Conference on Networks (ICON), 2006.

[28] B. Han, V. Gopalakrishnan, L. Ji, S. Lee, Network function virtualization: challenges and opportunities for innovations, Commun. Mag., IEEE 53 (2) (2015) 90–97.

[29] K. Gomez, L. Goratti, T. Rasheed, L. Reynaud, Enabling disaster-resilient 4g mobile communication networks, Commun. Mag., IEEE 52 (12) (2014) 66–73.

[30] M.L. Littman, Value-function reinforcement learning in Markov games, Cognit. Syst. Res. 2 (1) (2001) 55–66.

[31] A. Galindo-Serrano, L. Giupponi, Distributed Q-Learning for aggregated interference control in cognitive radio networks, IEEE Trans. Veh. Technol. 59 (4) (2010) 1823–1834.

[32] J.G. Schneider, W. keen Wong, A.W. Moore, M.A. Riedmiller, Distributed value functions, in: Proceedings of International Conference on Machine Learning (ICML), Bled, Slovenia, 1999.

[33] Y. Xing, R. Chandramouli, Stochastic learning solution for distributed discrete power control game in wireless data networks, IEEE Trans. Netw. 16 (4) (2008) 932–944.

[34] C.J.C.H. Watkins, Learning from delayed rewards., University of Cambridge, 1989 Ph.D. thesis.

[35] A.Y. Ng, S.J. Russell, Algorithms for inverse reinforcement learning, in: Proceedings of International Conference on Machine Learning (ICML), 2000, pp. 663–670.

[36] D. Broomhead, J. Huke, M. Muldoon, Codes for spread spectrum applications generated using chaotic dynamical system, Dyn. Stab. Syst. 14 (1) (1999) 95–105.

[37] H. Takagi, L. Kleinrock, Optimal transmission ranges for randomly distributed packet radio terminals, IEEE Trans. Commun. 32 (3) (1984) 246–257.

[38] H. Kim, K.G. Shin, In-band spectrum sensing in cognitive radio networks: energy detection or feature detection? in: Proceedings of the ACM Intl. Conference on Mobile computing and networking (MobiCom), San Francisco, CA, 2008.

[39] I.F. Akyildiz, W.-Y. Lee, K.R. Chowdhury, Crahns: cognitive radio ad hoc networks, Ad Hoc Netw. J. 7 (5) (2009) 810–836.

Eun Kyung Lee joined the PhD program at the Dept. of ECE, Rutgers University, in 2009. He is currently working under the guidance of Dr. Dario Pompili as a member of the NSF Center for Cloud and Autonomic Computing (CAC). His research interests include energy-efficient datacenter management and wireless sensor networks. Previously, he had received his BS in Electrical and Telecommunications Engineering from Soongsil University, South Korea and his MS in ECE from Rutgers University, in 2002 and 2009, respectively.

Hariharasudhan Viswanathan started his PhD program at the Dept. of Electrical and Computer Engineering (ECE), Rutgers University, in 2009. Currently, he is pursuing research in the fields of mobile computing, datacenter management, and wireless networking under the guidance of Dr. Dario Pompili at the NSF Center for Cloud and Autonomic Computing (CAC). Previously, he had received his BS in ECE from the PSG College of Technology, India and his MS in ECE from Rutgers University, in 2006 and 2009, respectively.

Dario Pompili is an Associate Professor with the Dept. of ECE at Rutgers University. He is the director of the Cyber Physical Systems Laboratory (CPS-Lab), which focuses on research problems in mobile computing, wireless communications and networking, sensor networks, and datacenter management. He received his PhD in ECE from the Georgia Institute of Technology in June 2007 under Dr. I. F. Akyildiz. He had previously received his 'Laurea' (integrated BS and MS) and Doctorate degrees in Telecommunications and System Engineering from the University of Rome "La Sapienza," Italy, in 2001 and 2004, respectively. He is a recipient of the prestigious NSF CAREER'11, ONR Young Investigator Program'12, and DARPA Young Faculty'12 awards.Since 2014, he is a Senior Member of both the IEEE Communications Society and the ACM.