IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received April 27, 2019, accepted May 15, 2019, date of publication June 4, 2019, date of current version June 19, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2920662

High-Reliability Multi-Agent Q-Learning-Based Scheduling for D2D Microgrid Communications

KEVIN SHIMOTAKAHARA, MEDHAT ELSAYED, KARIN HINZER, (Senior Member, IEEE), AND MELIKE EROL-KANTARCI, (Senior Member, IEEE)
School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON K1N 6N5, Canada

Corresponding author: Melike Erol-Kantarci (melike.erolkantarci@uottawa.ca)

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) under Grant 201703995, in part by the Collaborative Research and Training Experience Program (CREATE) under Grant 497981, and in part by the U.S. National Science Foundation (NSF) under Grant CNS-1647135.

ABSTRACT This paper proposes a multi-agent Q-learning-based resource allocation algorithm that allows long-term evolution (LTE)-enabled device-to-device (D2D) communication agents to generate the orthogonal transmission schedules outside the network coverage. This algorithm reduces packet drop rates (PDR) in distributed D2D communication networks to meet the quality-of-service requirements of the microgrid communications. The data traffic characteristics of three archetypal smart grid applications, namely demand response, solar, and generation forecasting, and synchrophasor communications, were simulated under seven different traffic congestion scenarios, where the total aggregate throughput of users ranged from 50% to 140% channel utilization. The PDR and latency performance of the proposed algorithm were compared with the existing random self-allocation mechanism introduced under the Third-Generation Partnership Project's LTE Release 12 standard for such scenarios. Our algorithm outperformed the LTE algorithm for all tested scenarios, demonstrating 20%–40% absolute reductions in PDR and 10–20-ms reductions in latency for all microgrid applications. The use of our algorithm in a simulated D2D-enabled demand response application resulted in a hundredfold reduction in power oscillations about the desired power flows.

INDEX TERMS Device-to-device communications, Q-learning, reliability, resource allocation, smart grid.

I. INTRODUCTION

To make smart grid solutions economically feasible, there is interest in implementing mobile wireless communication technology, i.e. Fourth Generation Long Term Evolution (4G LTE) and in the coming years Fifth Generation New Radio (5G NR) to establish a network of communication links over a distribution system with minimal investment in physical infrastructure [1]–[3]. Nevertheless, the data transfer delay (latency) in existing mobile communication technology is not guaranteed to be satisfactory for latency-critical smart grid services (e.g. synchrophasor applications), and addressing this problem is an ongoing area of research [3]–[5].

Device-to-Device (D2D) communication is a promising means to improve communication performance at a neighborhood area network scale by allowing direct data exchange between users, which in certain transmission modes can be controlled directly by such users [6], [7]. Moreover, such transmission modes also have the benefit of being usable even when devices are not within the coverage of a cell tower. Among the transmission modes that allow users to self-allocate resources are Transmission Mode 2 (TM-2) introduced in LTE Release 12, and TM-4 introduced in LTE Release 14. TM-2 was originally designed for public safety applications and prioritized prolonging battery life [8], but requires that D2D agents self-allocate cellular resources in a random manner and thus is not currently suitable for high reliability and low latency communications. On the other hand, TM-4 was designed for high reliability and low latency communication in addition to being able to manage fast-moving devices to facilitate vehicle-to-vehicle (V2V) communications, but disregards battery life considerations [8]. As such, neither transmission mode has been designed as an optimal smart grid communications solution. It can be argued that based on the communication requirements of smart grid applications, some combination of elements from both transmission modes would be best. We propose that the resource allocation mechanism of TM-2 be upgraded to meet the quality of service (QoS) requirements of smart grid applications in order to have a less complex and more power-efficient solution to D2D-enabled smart grid communications than TM-4.

The associate editor coordinating the review of this manuscript and approving it for publication was Pasquale De Meo.

A critical issue with TM-2 that impedes its communication performance is how it randomly self-allocates cellular resources, which leads to high packet drop rates (PDR), making it inadequate for certain smart grid applications sensitive to information losses. For example, a typical frequency stability synchrophasor application is sensitive to message failure rates exceeding 0.33% based off data loss sensitivity metrics published by the North American Synchrophasor Institute [9]. To overcome this problem, we propose a multi-agent high-reliability Q-learning (HRQ) resource allocation scheme to replace the random allocation mechanism. Naturally, HRQ is powered by Q-learning, which is a decision-making algorithm that considers the situation it is in, and chooses the best action to take based on past experience. Formally, that which takes the action is referred to as the ''agent'', and this agent interacts with its ''environment'' by taking ''actions''. The agent's situation is considered the ''state'' of the environment, and there are a finite set of actions that can be taken in any given state. The best action for a given state is the one with the largest expected ''reward'' associated with it, which is based on an action's influence on the environment [10]. HRQ divides the LTE resource grid into orthogonal partitions of resources from which a scheduling agent can select, and the agent is rewarded or penalized depending on whether or not it takes the same scheduling action as one or more other agent(s).

Concerning existing works on distributed resource allocation algorithms, few integrate the specifications of the LTE D2D TM-2 standard into their solution. Seemingly only Shih *et al.* [11] does so, offering a resource allocation algorithm to reduce packet drop rates. Our algorithm alternatively does not require randomly altering the operation of the D2D transmitter at the physical layer to sense control channel messages of other transmitting nodes in order to function. Other works either do not take into account the need to be implementable under LTE protocol architecture [12], have a centralized allocation scheme that is not applicable in out of coverage scenarios [13], [14], or focus instead on using D2D communications for frequency reuse in order to boost channel throughput instead of minimizing PDR for the D2D nodes [15]–[21].

The rest of the paper is organized as follows. Section II presents the system and traffic models, as well as the problem formulation. In Section III, the HRQ scheme is presented. Section IV contains the simulation results that show how the HRQ-enabled agents can achieve orthogonal self-organization of resources, causing lower latency and zero packet drop rate for low to moderate network traffic. Finally, Section V concludes the paper.

II. SYSTEM MODEL

A. MICROGRID COMMUNICATION NETWORK MODEL

Fig. 1a visualizes the communication network topology modeled in this study. We consider a set $\aleph$ of D2D nodes where each D2D node is $i \in \aleph$. A subset of D2D nodes $\Re \subset \aleph$ represent microgrid devices that have certain quality of service requirements. In addition, a set $\Im$ of auxiliary D2D nodes, where $\Im \subset \aleph$ are considered in order to overload the network with traffic and assess the performance of our algorithm. This auxiliary traffic is stochastic with message generation rates following a Poisson process, and message payloads are generated with exponential random variables. We refrain from modeling cellular nodes in our model, and instead consider an out of network coverage scenario. However, interference among D2D nodes remains due to the distributed and uncooperative resource allocation process.

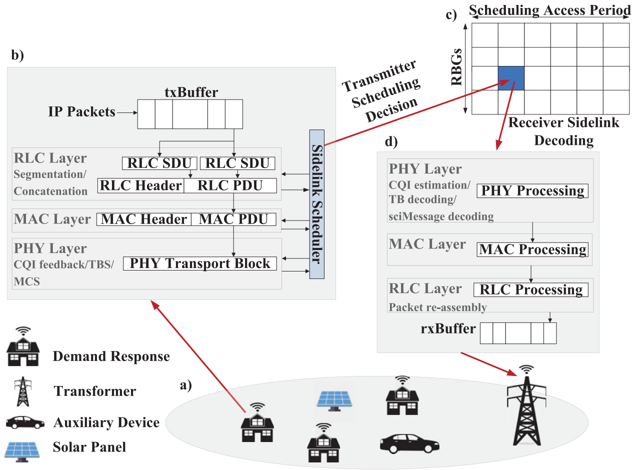

FIGURE 1. Network model of 3GPP LTE-compliant Release-12 D2D TM-2 communication. Concerning the acronyms in this figure, MAC is medium access control, PHY is physical, SDU is service data unit, PDU is protocol data unit, CQI is channel quality indicator, RBGs is resource block groups, MCS is modulation and coding scheme, tx is transmission, and rx is reception.

Fig. 1b-d depicts the data plane architecture of the LTE TM-2 standard used in this study. As can be seen in fig. 1b and fig. 1d, we model the lowest 3 layers of the LTE protocol stack. The radio link control (RLC) entities of the D2D devices have been modeled to operate in unacknowledged mode as per TM-2 specifications [22]. At the physical layer, TM-2 D2D communication channels, i.e. sidelink channels, are organized into repeating segments of the LTE resource grid known as scheduling assignment periods (fig. 1c). The scheduling assignment period architecture is further partitioned into a sidelink control channel, and a sidelink data channel. The sidelink control channel precedes the sidelink data channel in time, and D2D nodes make scheduling decisions once per scheduling assignment period, encoding a control message that the receiver can use to determine which resources to listen to in the upcoming sidelink data channel.

B. TRAFFIC MODEL OF THE MICROGRID APPLICATIONS

As presented in Table 1, we have modeled the traffic characteristics of three different microgrid applications, namely demand response, solar generation forecasting, and phasor management unit (PMU) communications. This table also includes some typical QoS requirements that could be expected for the aforementioned application archetypes. For the solar panel and demand response traffic, all parameters except their tolerable PDRs were derived based on the communication requirements specified by the US Department of Energy [4], [23]. For DR applications, we assume 1-9% PDR can be manageable based on Kong's analysis of dynamic pricing applications [24]. For distributed energy resource generation forecasting communication, it has been assumed that PDR should be close to 0 [25]. Concerning the PMU traffic model, Table 1 was populated using the PMU applications requirements published by the North American SynchroPhasor Institute (NAPSI) [9] in addition to the IEEE C37.118.2 standard for PMU-generated synchrophasor data transmission [26], [27].

TABLE 1. D2D Traffic settings and QoS requirements [4], [9], [23]–[25].

D2D application	Inter-arrival time [ms]	Packet size [b]	Max latency [ms]	Max PDR [%]
DR	20	2000	500-minutes	1-9
Solar	20	1120	300-2000	0
PMU	16	592	50	0.33
Auxiliary	20	2000	N/A	N/A

C. PROBLEM FORMULATION

The proposed solution aims to reduce the PDR of the D2D smart grid devices by performing efficient resource allocation in time and frequency. To improve reliability, signal-to-interference plus noise ratio (SINR) is used, where improving SINR improves the probability of successfully decoding the transmitted packets. SINR is formulated as follows:

$$\gamma_{u,k} = \frac{b_{u,k}p_{u,k}h_{u,k}}{\omega_k N_0 + \sum_{m\in I} b_{m,k}p_{m,k}h_{m,k}}, \tag{1}$$

where $\gamma_{u,k}$ is SINR of u^{th} D2D pair, and I is the set of interfering D2D pairs, i.e. the D2D pairs that use same resource blocks as transmitting u^{th} D2D pair. For the remaining terms, $b_{u,k}$ is allocation indicator of u^{th} D2D pair; $p_{u,k}$ is the transmit power of the u^{th} D2D pair on the k^{th} resource block; $h_{u,k}$ is the channel coefficient of k^{th} resource block; ω_k is bandwidth of k^{th} resource block; N_0 is additive white Gaussian noise (AWGN) single-sided power spectral density; $b_{m,k}$ is allocation indicator of m^{th} D2D interfering pair; $p_{m,k}$ is transmit power of interfering m^{th} pair; and $h_{m,k}$ is the channel coefficient of m^{th} interfering pair. Here, our optimization problem aims at maximizing the aggregate SINR of smart grid D2D pairs as follows:

$$\max_{b_{u,k}} \sum_{i=1}^{N}\sum_{k=1}^{K} \gamma_{u,k}2 \tag{2}$$

where N is the number of D2D pairs within the network, and K is the total number of resource blocks available. It should be noted that our objective function is solved subject to fixed power transmission. Moreover, the effects of device mobility are assumed negligible due to the assumption that the D2D nodes are stationary smart grid devices.

D. DEMAND RESPONSE ENABLED POWER SYSTEM MODEL

Smart grids are a combination of two sophisticated systems, namely a traditional power system and a communications network. To test the HRQ algorithm's impact on the performance of the communication network in relation to the LTE standard, it is only necessary to model the data traffic of the smart grid applications considered in this study. However, the demand response (DR) application was modeled on the power systems side as well in order to witness the impacts of communication system performance on the power system dynamics of a DR-enabled microgrid. The power system environment within which this application is being simulated can be visualized in fig. 2 [28]. The details of the simulation environment is provided in Section IV. In our model, if the power flow through a transformer crosses a certain threshold level P_{limit}, a "demand response event" is initiated. If such an event happens, the smart meter continuously recalculates the minimum demand reductions required to keep the power consumption from exceeding P_{limit}, and sends messages to the households containing new power reduction requests. Also, households were set to reduce their power consumption to a constant "base comfort" level as long as the price of electricity surpasses a particular value C_{max}. It is assumed that households have instantaneous control over their power consumption.

III. HIGH-RELIABILITY Q-LEARNING (HRQ) SCHEME

HRQ is a multi-agent distributed Q-Learning algorithm performed by each D2D transmitter to efficiently allocate resource blocks every scheduling assignment period for reliability maximization. HRQ models reliability in terms of SINR of D2D devices as per equation (1), as low SINR is a root cause of PDR. In particular, efficient resource block allocation allows devices to select disjoint actions, i.e. distinct resource blocks, which lessens interference and increases SINR. As such, improving SINR increases the probability of successfully decoded packets which achieves higher reliability. Besides reliability, interference mitigation enables devices to transmit on channels of higher quality that lead to allocation of higher transport block size (TBS). As such, large packets are less prone to segmentation at the radio link control layer, leading to reduced latency. It should be noted that this algorithm manages the allocation of resource blocks, i.e. physical spectrum, and is not designed to manage transmitter power settings.

In HRQ, we address the reliability by formulating the Q-Learning tuple as follows:

- **Agents:** D2D transmitting nodes.
- **States:** Channel quality is used to represent HRQ states, where we define CQI_{ideal} as the best channel quality that drives devices to allocate disjoint resources, hence achieving highest reliability. In simulations, we define stringent channel quality requirements as CQI_{ideal}.

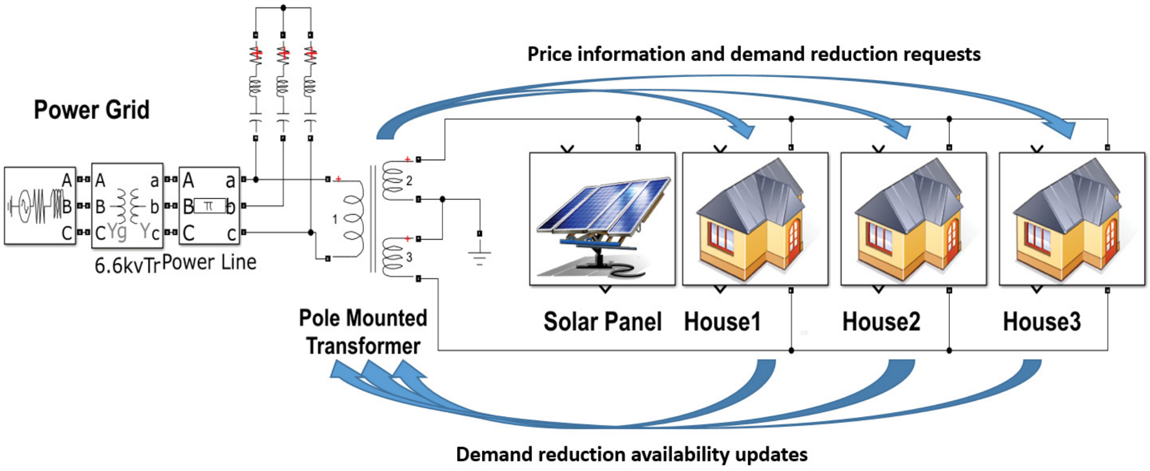

FIGURE 2. Diagram of power system modeled in the integrated simulator.

- **Actions:** HRQ performs actions that manifest as the selection of cellular resources in time and frequency every scheduling assignment period. To minimize the action space, contiguous sets of subframes and resource blocks are aggregated to form larger (but still orthogonal) sets of resources that agents can select as their scheduling decision. In frequency, every four contiguous resource blocks are combined to form a ''resource block group'', whereas in time, the pool is partitioned into two 16 subframe intervals. The network model was set to a 5 MHz bandwidth and 40 ms scheduling assignment period, which consists of 25 resource blocks and 32 subframes for data transmission. Under such conditions, 12 actions populate the action space.
- **Reward function:** The reward is defined as follows:

$$\tau_u = \begin{cases} ll{-}1 & CQI_u < CQI_{ideal}, \\ 1 & \text{otherwise}, \end{cases} \tag{3}$$

where τ_u is the reward of u^{th} D2D pair, CQI_u is the channel quality of u^{th} D2D pair, and CQI_{ideal} is the ideal channel quality. The rationale behind the reward function is to reward the agent when achieving the ideal QoS, i.e. ideal channel quality, whereas it penalizes the agent otherwise.
- **Q-value update:** The Q-values are updated according to the temporal difference (TD) equation [10]:

$$\begin{aligned} QV\,(S_u, a_u) &= (1-\alpha)QV\,(S_u, a_u) \\ &\quad + \alpha[\tau_u + \gamma \max_{A_u} QV\left(S'_u, A_u\right)] \end{aligned} \tag{4}$$

where α is a learning rate, γ is a discount factor, A_u is the action-space of u^{th} D2D pair, and $QV(S'_u, A_u)$ are the Q-values at next state S'_u and all actions A_u. The new action a'_u is selected based on a modified version of ϵ-greedy approach:

$$a'_u = \begin{cases} random & \epsilon, \\ random(arg \max_{A_u} QV(S'_u, A_u)) & 1-\epsilon, \end{cases} \tag{5}$$

where ϵ is the exploration probability. The modified ϵ-greedy accounts for situations where multiple state-action pairs are tried for having the largest Q-value during exploitation iterations. Thus, during the exploitation phase, the scheduler chooses among the set of actions with the highest Q-value at random instead of the conventional approach of selecting the first Q value in the list returned by a max function.

Algorithm 1 HRQ

```
Initialization: Q-table $\leftarrow 0, \alpha, \gamma$, and $\epsilon$.
for scheduling assignment period $t = 1$ to $T$ do
    Step 1: Update channel quality value based on the last transmission.
    Step 2: Compute the reward $\tau_u$ and observe the new state $S'_u$ due to last action execution.
    Step 3: Update the Q-value as in (4).
    Step 4: Transit to next state $S'_u$.
    Step 5: Select the next action $a'_u$ based on modified $\epsilon$-greedy policy as in (5).
    If Exploitation and tie then
    Select action Randomly
    End If
end for
```

Algorithm 1 presents the steps of HRQ. TM-2 currently has no means to feed back channel quality metrics to the transmitters, but a new sidelink control information message that can be sent by the receiver back to the sender would be a simple solution to this. The algorithm terminates after T scheduling assignment periods.

This study compares the performance of the proposed HRQ scheduling algorithm to that of the allocation scheme prescribed by the LTE standard for D2D communication operating in TM-2 [29]. The step-by-step process of our implementation of the TM-2 random self-allocation algorithm have been included in the Appendix.

IV. RESULTS

This section is divided into three subsections. Subsection IV-A presents our results on the latency and PDR performance of both scheduling algorithms, where 20-40% reductions in PDR and >10 ms drops in latency were observed for all smart grid applications. Subsection IV-B shows the convergence of the HRQ algorithm; HRQ was able to converge as long as the number of agents did not exceed the size of the action space. Subsection IV-C presents the results collected on the observed changes in power system dynamics of the simulated DR-enabled microgrid as a function of resource allocation mechanism selection. The power fluctuations observed were reduced by two orders of magnitude when HRQ was used in lieu of random self-allocation.

For our simulations, we implemented the D2D communication protocol stack on top of MATLAB's LTE Toolbox which was then combined with a microgrid simulator implemented using Simscape Power Systems and SimEvents software packages available in the Matlab/Simulink environment. The performance and convergence of the HRQ algorithm was tested using the mobile communications portion of the developed simulator, whereas all power and communication elements were leveraged in simulating the DR-enabled microgrid. The simulation parameters are given in Table 2 in the Appendix.

A. COMPARING THE PERFORMANCE OF HRQ AND LTE TM-2 SCHEDULING STRATEGIES UNDER VARYING TRAFFIC INTENSITIES

Under the same network and traffic conditions, the HRQ and LTE schedulers were tested for various levels of network traffic. The network traffic was adjusted by how many "auxiliary" D2D devices were injected into the communication network. At each level of network traffic tested (i.e. for 0, 4, 5, 6, 7, 9, and 11 auxiliary devices present in the network), the communication network was simulated for 15 s, over which time 375 scheduling assignment periods (and thus scheduling decisions) elapsed. Moreover, for every 15 s simulation at every level of network traffic, the simulation was repeated 12 times, and the average message latency and PDR values were plotted with 95% confidence intervals for both scheduling strategies.

Over these tests, the exploration time of the HRQ algorithm was set to 10 s, during which time an epsilon greedy exploration strategy was implemented by the scheduler. After 10 s, the HRQ algorithm switched to a purely greedy strategy, only executing its policy without exploration. This test compared the performance of HRQ with respect to LTE scheduling strategies after HRQ has completed its exploration phase. Thus, the performance metrics obtained are calculated based only on the final 5 s of the network simulation.

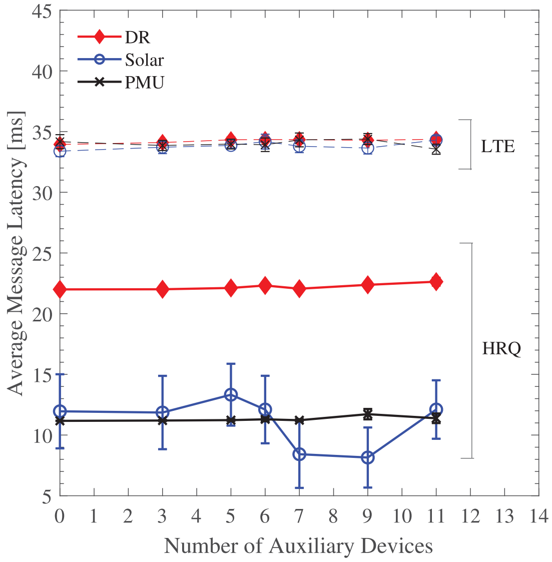

FIGURE 3. Average message latency for smart grid applications under varying traffic intensity.

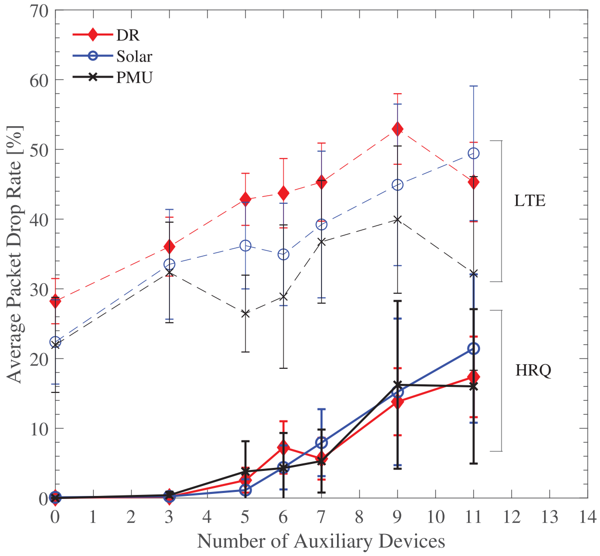

FIGURE 4. Average PDR for smart grid applications under varying traffic intensity.

The results of latency and PDR performance for both schedulers have been presented in fig.s 3 and 4 respectively. Fig. 3 demonstrates the potential for D2D communication to provide strong latency performance; both scheduling options succeed in meeting the QoS requirements of the smart grid applications, which is due to the inherent nature of D2D communication. However, HRQ provides additional improvements in latency, as it proactively allocates communication resources every scheduling assignment period instead of scheduling resources for what has been buffered since the last scheduling assignment period.

Concerning PDR performance, it can be seen in fig. 4 that for varying traffic intensities, the LTE scheduler fails to meet the QoS requirements for the smart grid applications. However, the PDR can be kept down to essentially 0 for low to moderate levels of network traffic when using the HRQ scheduler. The PDR starts to rise as the number of scheduling agents in the network approaches the number of orthogonal scheduling decisions available in the action space. The number of agents in the network reaches the number of orthogonal scheduling decisions when 6 auxiliary devices are injected in the network. This results from the inability of the HRQ algorithm to consistently self-organize the agent's scheduling decisions in the 10 s exploration time provided in this experiment, which is further discussed in Section IV-B. Naturally, the average PDR and uncertainty around the average PDR continues to increase as the number of agents in the network begin to exceed the number of actions available for the agents to take, as it becomes impossible to guarantee QoS with HRQ under such conditions. For example, if there are 12 D2D pairs each taking a unique action among the 12 available, if another D2D pair is introduced to the network, any action it chooses will be the same as what is already being taken.

B. CONVERGENCE OF HRQ ALGORITHM

To demonstrate the ability for HRQ to reach an optimal policy, which corresponds to all agents making unique (and orthogonal) scheduling decisions, cumulative regret of all smart grid agents were aggregated and plotted with respect to time in fig. 5. Regret is the difference between the reward of an action taken and the reward associated with the action that an optimal policy would have taken. Cumulative regret is the time integral of regret. Because the HRQ scheduler reward function has only two possible outputs, namely 1 and −1, every time an agent takes an action that results in a −1 (which happens when an agent makes the same scheduling decision as another agent), the regret is 2. Otherwise, the regret is 0. Thus, the point in the simulation where the smart grid agents are capable of self organizing their scheduling decisions so that they do not interfere with one another corresponds to the point in time where the cumulative regret of all agents stops increasing.

It can be observed in that the HRQ algorithm is capable of converging to an optimal policy when the number of agents does not exceed the number of available orthogonal scheduling decisions in the algorithm's action space. When 6 auxiliary devices are introduced to the network, the number of agents in the network matches the number of actions defined in the HRQ action space. In this case, the HRQ algorithm takes over 15 s to converge (recall that one subframe is 1 ms). This is why the average PDR was larger than 0% in fig. 4 even in cases where there were a sufficient number of actions for each agent to assign themselves a unique action. When 9 auxiliary devices are introduced to the network, fig. 5 results suggest that a plateau in cumulative regret is about to be formed at the very end of the time axis. This is possible because the aggregate cumulative regret of the smart grid agents plotted does not include the cumulative regrets of the auxiliary agents. To elaborate, there is the possibility that the set of smart grid agents could all terminally select disjoint actions while the auxiliary agent's HRQ policies converge to taking actions (that are not all disjoint) among the remainder of the action space.

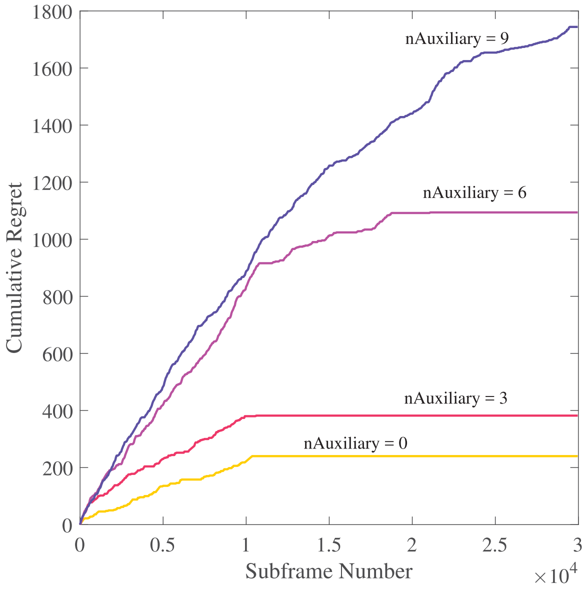

FIGURE 5. Total cumulative regret for HRQ scheduling algorithm for various numbers of auxiliary devices.

C. IMPACT OF SCHEDULING TECHNIQUE ON THE PERFORMANCE OF MICROGRID APPLICATIONS

Fig. 6 demonstrates the functionality of the application performing demand response activity due to both price signals and transformer overloading scenarios as described in Section II-D. Fig. 6a plots the aggregate power curves of the households, and fig. 6b plots the transformer's power curves for multiple cases. The first case is where the demand response application is disabled (no DR), providing a baseline power curve that can be altered by the application. The next three cases correspond to the dynamics of the power system when demand response is active and facilitated by D2D LTE communications utilizing the LTE scheduling mechanism, HRQ, and an ideal scheduler (ideal data transfer). The ideal scheduler allows the smart grid devices to communicate under perfect channel conditions that ignore the effects of interference and noise. The power curve for the photovoltaic (PV) module and the real time price of electricity are plotted in fig. 6c and fig. 6d respectively. The transformer power flow curve is the difference of household aggregate demand less PV power generation. In cases where

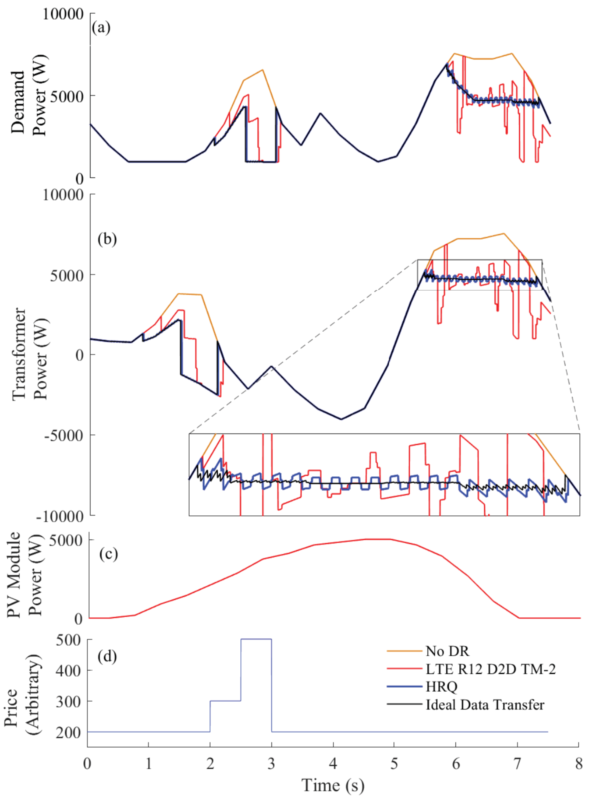

FIGURE 6. Demand response application performance for 3GPP, HRQ, and ideal mobile communication models: Plots of a) Aggregate household demand; b) transformer power flow; c) microgrid PV generation; and d) price of electricity as a function of time.

transformer power flow is positive, the distribution system is providing power to the microgrid; when it is negative, the power generated by the PV exceeds the aggregate household demand, and the surplus flows into the distribution system.

Concerning the price signaling dynamics of the three homes in our model (fig. 2), one home is programmed to have a C_{max} of 300 price units, whereas the other two homes are given C_{max} values of 400 price units. Based on this, one should expect power consumption to drop when the price curve increases, and fall back down again when the price decreases. Also, the transformer power flow exceeds $P_{limit} = 5000W$ shortly before the 6 s mark of the simulation, meaning the demand should decrease to keep the power flow through the transformer at or below $5000W$. As can be observed, the ideal data transfer and HRQ power curves do manage to behave in this manner, and are almost identical to one another. However, the power curve in the scenario where the application is being implemented with the LTE resource scheduling algorithm proves to deviate from the ideal case. For example, it can be seen at the 2 and 2.5 second marks, there are delays in the responsiveness of the households on the order of 100 ms when they are reacting to price changes. Such delays are not present in the other curves. This can only be a consequence of the higher PDR present when utilizing the LTE scheduler. These plots are intriguing in the sense that they display the ability of the integrated simulation platform to investigate how the performance of the communication network influences that of the power system.

V. CONCLUSION

This paper proposed a multi-agent Q-Learning based resource allocation strategy targeted at reducing packet drop rates in LTE TM-2 D2D communication to increase its usability for smart grid applications. The QoS requirements and data traffic characteristics for various smart grid application archetypes were studied and modeled in an integrated Matlab/Simulink simulation environment to both test the performance of the scheduling mechanism and demonstrate the simulator's ability to design smart power systems and observe how the power and communications systems interact with one another. The performance of the HRQ scheduling algorithm was compared to that of the existing means of resource scheduling in TM-2 prescribed by the LTE standard. Our results show that HRQ outperformed the benchmark algorithm in both latency and PDR QoS metrics. Finally, a basic demand response application developed in the simulator was presented to demonstrate the impact the two different scheduling strategies have on its ability to properly manage transformer loading and price signaling activities. As a future work, we plan to enhance the performance of HRQ with advanced exploration strategies and implement more sophisticated smart grid use cases.

APPENDIX A
RESOURCE ALLOCATION ALGORITHM AND SIMULATION IMPLEMENTATION DETAILS

A. LTE RANDOM ALLOCATION ALGORITHM

TS 36.321 Release 12 Section 5.14.1.1 specifies to randomly select the time and frequency resources for sidelink shared channel (SL-SCH) and sidelink control information of a sidelink grant from the resource pool configured by upper layers with equal probability [29]. This specification is implemented as **Algorithm 2** in our simulator.

Conceptually, the scheduling algorithm calculates how much data is queued in the device's transmission buffer, and then determines all feasible combinations of I_{trp} (time resource pattern index) and RIV (Resource Indication Value) sidelink control information scheduling parameters that correspond to the smallest amount of bandwidth capable of sending all the existing data queued over the course of the next scheduling assignment period. Finally, the scheduler randomly selects among all feasible scheduling options that were found this way with equal probability.

The variables of **Algorithm 2** have been defined as follows:

- $TBSdesired$: The TBS that would result in the scheduler being able to clear the data buffer of the agent over the course of the next scheduling assignment period. It should be noted that any data that arrives after the scheduling decision must wait until the next scheduling decision in order to be allocated resources.

Algorithm 2 Implementation of 3GPP Release 12 D2D TM-2 Random Allocation Scheduling

```
Initialization: $nprbMax, I_{TBS}$.
for scheduling assignment period $t = 1$ to $T$ do
    $k = 1$
    for $nTBs$ = Each possible number of transport blocks to transmit next scheduling assignment period do
        Step 1: Compute $TBSdesired$.
        Step 2: Iterate through TS 36.213 Release V12.5.0 Section 7.1.7.2 Table 7.1.7.2.1-1 until it is discovered how many PRBs are required to generate a transport block with a TBS equal to or greater than $TBSdesired$:
        for $i = 1$ to $nprbMax$ do
            Read the TBS value corresponding to $I_{TBS}$ and $i$ PRBs.
            If TBS value read >= $TBSdesired$
            $nPRBs \leftarrow i$.
            $success \leftarrow$ true
            Break from innermost loop
            end If
        end for
        If $success$
        $nTBnPRBcombos(k) \leftarrow [nTBs, nPRBs]$
        $k \leftarrow k + 1$
        end If
    end for
    Step 3: For every set $[nTBs, nPRBs]$ in $nTBnPRBcombos(k)$, determine all $[RIV, I_{TRP}]$ pairs possible, and out of ALL pairs across ALL $[nTBs, nPRBs]$ sets, choose one at random as the final scheduling decision for the scheduling assignment period.
end for
```

TABLE 2. Network settings.

Network Settings	
Transmission bandwidth	5 MHz
Number of resource blocks	25 (12 subcarriers / resource block)
Scheduling assignment period	40 msec
Time-to-transmit interval (TTI)	1 msec
D2D resource block pool	25
D2D subframe pool index range	8-39
Microgrid radius	50 m
Number of microgrid D2D nodes	7
Number of auxiliary D2D	0-11
Channel Settings	
Transmit power	20 dBm
Tx/Rx antenna gain	10 dB
Pathloss	3GPP pathloss model [30]
Penetration loss	5 dB
Noise Figure	5 dB
Shadowing	$\sim$ LOGN(0, 2(dB))
Q-Learning Settings	
Learning rate (α)	0.5
Discount factor (γ)	0.9
Exploration probability (ϵ)	0.1

- $nprbMax$: The largest number of PRBs a scheduler can self-allocate given the bandwidth of the scheduling assignment period. In this study, a 5MHz scheduling assignment period was used, so $nprbMax$ was set to 25 PRBs.
- I_{TBS}: An index that maps a particular MCS (Modulation and Coding Scheme) to a row in an LTE lookup table (for example TS 36.213 Release V12.5.0 Section 7.1.7.2 Table 7.1.7.2.1-1) that ultimately yields TBS values for given combinations of MCS settings and the number of PRBs being used to map the TB to the resource grid.
- $nTBs$: A feasible number of transport blocks that can be sent by an agent over the course of a scheduling assignment period. For the given network configuration in this study, $nTBs$ can equal 1, 2, 4, or 8.
- $nPRBs$: The number of PRBs required to be allocated in order to have access to TBs with a size equal to or greater than $TBSdesired$ given a particular $nTBs$ value.
- $nTBnPRBcombos$: A list of all feasible $[nTBs, nPRBs]$ combinations that yield enough bandwidth to clear the agent's data buffer.
- RIV: A "Resource Indication Value" parameter specified by the LTE standard for D2D communications that specifies both the number of PRBs to be scheduled in the next scheduling assignment period in addition to which set of contiguous PRBs to use.
- I_{TRP}: A Time Resource Pattern Index specified by the LTE standard for D2D communications that specifies the number of subframes and their indices in the next scheduling assignment period. This ultimately controls how many TBs can be sent in a single scheduling assignment period by an agent and when.

B. COMPLEXITY ANALYSIS

For the sake of curiosity, we investigate the complexity of the LTE random allocation compared to the proposed Q-learning algorithm, HRQ. For a specific bandwidth configuration, the number of resource block groups is defined as $nRBGs = \lfloor \frac{nPRBs}{sRBG} \rfloor$, where $sRBG$ is the size of a resource block group in resource blocks. As shown in **Algorithm 2**, random allocation performs a search in Table 7.1.7.2.1-1 to find the required number of resource blocks to allocate. This search is performed four times for each possible value of $nTBs$. Assuming linear search, each possible $nTBs$ requires $nRBGs$ operations, hence a total number of $(4\, nRBGs)$ operations are needed. Therefore, the Big-O complexity of random allocation is $O(N)$, where $N = (4nRBGs)$.

Under same assumption of linear search, complexity of HRQ relies mainly on the maximum search performed in eq. (4) and (5). Therefore, complexity of HRQ can be identified using number of actions at a desired state, i.e. a row in the Q-table. Number of actions is determined by considering the possible allocation units in time and frequency direction which are $nTBs$ and $nRBGs$ respectively, hence total number of actions becomes $(nTBsnRBGs)$. As such, the

Big-O complexity of HRQ is also $O(N)$, but in this case $N = (nTBsnRBGs)$.

C. NETWORK SIMULATION SETTINGS

See Table 2.

ACKNOWLEDGMENT

The authors thank Dr. Javad Fattahi for his feedback on the initial draft of this paper.

REFERENCES

[1] X. Fang, S. Misra, G. Xue, and D. Yang, "Smart grid—The new and improved power grid: A survey," *IEEE Commun. Surveys Tuts.*, vol. 14, no. 4, pp. 944–980, 4th Quart., 2012.

[2] M. Erol-Kantarci and H. T. Mouftah, "Energy-efficient information and communication infrastructures in the smart grid: A survey on interactions and open issues," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 179–197, 1st Quart., 2015.

[3] I. Al-Anbagi, M. Erol-Kantarci, and H. T. Mouftah, "Delay critical smart grid applications and adaptive QoS provisioning," *IEEE Access*, vol. 3, pp. 1367–1378, 2015.

[4] V. C. Gungor, D. Sahin, T. Kocak, S. Ergut, C. Buccella, C. Cecati, and G. P. Hancke, "A survey on smart grid potential applications and communication requirements," *IEEE Trans. Ind. Informat.*, vol. 9, no. 1, pp. 28–42, Feb. 2013.

[5] F. A. Asuhaimi, J. P. B. Nadas, and M. A. Imran, "Delay-optimal mode selection in device-to-device communications for smart grid," in *Proc. IEEE Int. Conf. Smart Grid Commun. (SmartGridComm)*, Oct. 2017, pp. 26–31.

[6] C. Kalalas, L. Thrybom, and J. Alonso-Zarate, "Cellular communications for smart grid neighborhood area networks: A survey," *IEEE Access*, vol. 4, pp. 1469–1493, 2016.

[7] K. Shamganth and M. J. N. Sibley, "A survey on relay selection in cooperative device-to-device (D2D) communication for 5G cellular networks," in *Proc. Int. Conf. Energy, Commun., Data Anal. Soft Comput. (ICECDS)*, Aug. 2017, pp. 42–46.

[8] R. Molina-Masegosa and J. Gozalvez, "LTE-V for sidelink 5G V2X vehicular communications: A new 5G technology for short-range vehicle-to-everything communications," *IEEE Veh. Technol. Mag.*, vol. 12, no. 4, pp. 30–39, Dec. 2017.

[9] NASPI, "PMU data quality: A framework for the attributes of PMU data quality and a methodology for examining data quality impacts to synchrophasor applications, version 1.0," North American SynchroPhasor Initiative, Kauai, HI, USA, White Paper NASPI-2017-TR-002 PNNL-26313, Mar. 2017.

[10] E. Alpaydin, *Introduction to Machine Learning*. Cambridge, MA, USA: MIT Press, 2014.

[11] M.-J. Shih, H.-H. Liu, W.-D. Shen, and H.-Y. Wei, "UE autonomous resource selection for D2D communications: Explicit vs. implicit approaches," in *Proc. IEEE Conf. Standards Commun. Netw. (CSCN)*, Oct./Nov. 2016, pp. 1–6.

[12] Y. Cao, T. Jiang, M. He, and J. Zhang, "Device-to-device communications for energy management: A smart grid case," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 1, pp. 190–201, Jan. 2016.

[13] S. Wen, X. Zhu, X. Zhang, and D. Yang, "QoS-aware mode selection and resource allocation scheme for device-to-device (D2D) communication in cellular networks," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC)*, Jun. 2013, pp. 101–105.

[14] S. Alwan, I. Fajjari, and N. Aitsaadi, "Joint routing and wireless resource allocation in multihop LTE-D2D communications," in *Proc. IEEE 43rd Conf. Local Comput. Netw. (LCN)*, Oct. 2018, pp. 167–174.

[15] S. Maghsudi and S. Stańczak, "Joint channel allocation and power control for underlay D2D transmission," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2015, pp. 2091–2096.

[16] H. Ye and G. Y. Li, "Deep reinforcement learning for resource allocation in V2V communications," *CoRR*, Nov. 2017, pp. 1–6. [Online]. Available: http://arxiv.org/abs/1711.00968

[17] A. Asheralieva and Y. Miyanaga, "An autonomous learning-based algorithm for joint channel and power level selection by D2D pairs in heterogeneous cellular networks," *IEEE Trans. Commun.*, vol. 64, no. 9, pp. 3996–4012, Sep. 2016.

[18] S. Nie, Z. Fan, M. Zhao, X. Gu, and L. Zhang, "Q-learning based power control algorithm for D2D communication," in *Proc. IEEE 27th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Sep. 2016, pp. 1–6.

[19] F. A. Asuhaimi, S. Bu, and M. A. Imran, "Joint resource allocation and power control in heterogeneous cellular networks for smart grids," in *Proc. IEEE GLOBECOM*, Dec. 2018, pp. 1–6.

[20] A. Laya, K. Wang, A. A. Widaa, J. Alonso-Zarate, J. Markendahl, and L. Alonso, "Device-to-device communications and small cells: Enabling spectrum reuse for dense networks," *IEEE Wireless Commun.*, vol. 21, no. 4, pp. 98–105, Aug. 2014.

[21] L. Melki, S. Najeh, and H. Besbes, "Radio resource allocation scheme for intra-inter-cell D2D communications in LTE-A," in *Proc. IEEE 26th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Aug./Sep. 2015, pp. 1515–1519.

[22] Rohde and Shwarz, "1MA264: Device to device communication in LTE, version 0e," Rohde & Shwarz, Munich, Germany, White Paper 1MA264, Sep. 2015.

[23] *Communications Requirements of Smart Grid Technologies*, U.S. Dept. Energy, Washington, DC, USA, Jan. 2012.

[24] P.-Y. Kong, "Effects of communication network performance on dynamic pricing in smart power grid," *IEEE Syst. J.*, vol. 8, no. 2, pp. 533–541, Jun. 2014.

[25] R. A. Cacheda, D. C. García, A. Cuevas, F. J. G. Castaño, J. H. Sánchez, G. Koltsidas, V. Mancuso, J. I. M. Novella, S. Oh, and A. Pantò, "QoS requirements for multimedia services," in *Resource Management in Satellite Networks*. New York, NY, USA: Springer, Jan. 2007, pp. 67–94.

[26] *IEEE Standard for Synchrophasor Data Transfer for Power Systems*, Standard C37.118.2-2011, Dec. 2011.

[27] S. R. Firouzi, L. Vanfretti, A. Ruiz-Alvarez, F. Mahmood, H. Hooshyar, and I. Cairo, "An IEC 61850-90-5 gateway for IEEE C37.118.2 synchrophasor data transfer," in *Proc. IEEE Power Energy Soc. General Meeting (PESGM)*, Jul. 2016, pp. 1–5.

[28] *Simplified Model of a Small Scale Micro-Grid—MATLAB & Simulink*. Accessed: Feb. 7, 2019. [Online]. Available: https://www.mathworks.com/help/physmod/sps/examples/simplified-model-of-a-small-scale-micro-grid.html

[29] *Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) Protocol Specification, Version 12.5.0*, document 136.321, 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 3GPP, Apr. 2015.

[30] C. C. Coskun and E. Ayanoglu, "Energy- and spectral-efficient resource allocation algorithm for heterogeneous networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 1, pp. 590–603, Jan. 2018.

KEVIN SHIMOTAKAHARA received the B.Eng. degree from Carleton University, in 2018. He is currently pursuing the M.A.Sc. degree with the University of Ottawa. His research interests include smart grids, renewable energy, and artificial intelligence.

MEDHAT ELSAYED received the B.Sc. and M.Sc. degrees from Cairo University, Egypt, in 2009 and 2013, respectively. He is currently pursuing the Ph.D. degree with the University of Ottawa. His research interests include AI-enabled wireless networks, 5G and beyond, and smart grids.

KARIN HINZER received the B.Sc., M.Sc., and Ph.D. degrees in physics from the University of Ottawa, Ottawa, ON, Canada, in 1996, 1998, and 2002, respectively.

She was with the National Research Council Canada, Nortel Networks, and Bookham (now Oclaro), where she gained extensive experience in the design and fabrication of the group III–V semiconductor devices. Cost reduction strategies and liaison with remote fabrication facilities strongly feature in her industry experience. In 2007, she joined the University of Ottawa, where she founded the SUNLAB, the premier Canadian modeling and characterization laboratory for next-generation multi-junction solar devices and concentrator systems. Her research involves developing new ways to harness the solar energy. From 2007 to 2017, she was the Tier II Canada Research Chair in Photonic Nanostructures and Integrated Devices. Her laboratory has spun off three Canadian companies in the energy sector. She has published over 160 refereed papers and trained over 110 highly qualified personnels. Her research interests include new materials, high-efficiency light sources and light detectors, solar cells, solar modules, new electrical grid architectures/controls, and voltage converters. She is currently a Professor with the School of Electrical Engineering and Computer Science with a cross-appointment at the Department of Physics, University of Ottawa.

Dr. Hinzer is also a member of the College of New Scholars, Artists, and Scientists of the Royal Society of Canada. In 2010, she was a recipient of the Inaugural Canadian Energy Award with industry partner Morgan Solar for the development of more efficient solar panels. In 2015, she received the Ontario Ministry of Research and Innovation Early Researcher Award for her contributions to the fields of photonic devices and photovoltaic systems. In 2016, she was a recipient of the University of Ottawa Young Researcher Award. She is also the Principal Investigator of the Natural Sciences and Engineering Research Council of Canada Collaborative Research and Training Experience Program titled Training in Optoelectronics for Power: From Science and Engineering to Technology (NSERC CREATE TOP-SET), a multi-disciplinary training program involving three universities and training over 100 students in six years. She is also an Editor of the IEEE Journal of Photovoltaics.

MELIKE EROL-KANTARCI received the M.Sc. and Ph.D. degrees in computer engineering from Istanbul Technical University, in 2004 and 2009, respectively. During her Ph.D. studies, she was a Fulbright Visiting Researcher with the Computer Science Department, University of California at Los Angeles (UCLA). She is currently an Associate Professor with the School of Electrical Engineering and Computer Science, University of Ottawa. She is also the Founding Director of the Networked Systems and Communications Research (NETCORE) Laboratory. She is also a courtesy Faculty Member with the Department of Electrical and Computer Engineering, Clarkson University, Potsdam, NY, USA, where she was a tenure-track Assistant Professor prior to joining the University of Ottawa. She has over 100 peer-reviewed publications that have been cited over 3900 times. She has an h-index of 30. She is the Co-Editor of two books: *Smart Grid: Networking, Data Management, and Business Models* (CRC Press) and *Transportation and Power Grid in Smart Cities: Communication Networks and Services* (Wiley). Her main research interests include AI-enabled networks, 5G and beyond wireless networks, smart grid, electric vehicles, and the Internet of Things. She received the IEEE Communication Society Best Tutorial Paper Award and the Best Editor Award of the IEEE Multimedia Communications Technical Committee, in 2017. She has acted as the General Chair or the Technical Program Chair for many international conferences and workshops. She was also the past Vice-Chair for Women in Engineering (WIE) at the IEEE Ottawa Section. She is also the Chair of Green Smart Grid Communications special interest group of the IEEE Technical Committee on Green Communications and Computing. She is also an Editor of the IEEE Communications Letters and IEEE Access.