Energy-Efficiency Oriented Traffic Offloading in Wireless Networks: A Brief Survey and a Learning Approach for Heterogeneous Cellular Networks

Xianfu Chen, *Member, IEEE*, Jinsong Wu, *Senior Member, IEEE*, Yueming Cai, *Senior Member, IEEE*, Honggang Zhang, *Senior Member, IEEE*, and Tao Chen, *Senior Member, IEEE*

***Abstract*—This paper first provides a brief survey on existing traffic offloading techniques in wireless networks. Particularly as a case study, we put forward an online reinforcement learning framework for the problem of traffic offloading in a stochastic heterogeneous cellular network (HCN), where the time-varying traffic in the network can be offloaded to nearby small cells. Our aim is to minimize the total discounted energy consumption of the HCN while maintaining the quality-of-service (QoS) experienced by mobile users. For each cell (i.e., a macro cell or a small cell), the energy consumption is determined by its system load, which is coupled with system loads in other cells due to the sharing over a common frequency band. We model the energy-aware traffic offloading problem in such HCNs as a discrete-time Markov decision process (DTMDP). Based on the traffic observations and the traffic offloading operations, the network controller gradually optimizes the traffic offloading strategy with no prior knowledge of the DTMDP statistics. Such a model-free learning framework is important, particularly when the state space is huge. In order to solve the curse of dimensionality, we design a centralized Q-learning with compact state representation algorithm, which is named QC-learning. Moreover, a decentralized version of the QC-learning is developed based on the fact the macro base stations (BSs) can independently manage the operations of local small-cell BSs through making use of the global network state information obtained from the network controller. Simulations are conducted to show the effectiveness of the derived centralized and decentralized QC-learning algorithms in balancing the tradeoff between energy saving and QoS satisfaction.**

***Index Terms*—Wireless networks, heterogeneous cellular networks, traffic offloading, energy saving, traffic load balancing, discrete-time Markov decision process, reinforcement learning, compact state representation, team Markov game.**

Manuscript received July 13, 2014; revised September 18, 2014 and December 16, 2014; accepted December 24, 2014. Date of publication January 16, 2015; date of current version April 17, 2015. This work was supported in part by the National Basic Research Program of China (973Green, No. 2012CB316000), by the Key (Key grant) Project of Chinese Ministry of Education (No. 313053), by the Key Technologies R&D Program of China (No. 2012BAH75F01), and the grant of "Investing for the Future" program of France ANR to the CominLabs excellence laboratory (ANR-10-LABX-07-01).

X. Chen and T. Chen are with VTT Technical Research Centre of Finland Ltd., Oulu 90570, Finland (e-mail: xianfu.chen@vtt.fi; tao.chen@vtt.fi).

J. Wu is with the Department of Electrical Engineering, Universidad de Chile, Santiago 1058, Chile (e-mail: wujs@ieee.org).

Y. Cai is with PLA University of Science and Technology, Nanjing 210007, China (e-mail: caiym@vip.sina.com).

H. Zhang is with Zhejiang University, Hangzhou 310027, China. He was also with the Université Européenne de Bretagne (UEB), Rennes 35000, France and the Supélec, Cesson-Sévigné 35576, France (e-mail: honggangzhang@zju.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSAC.2015.2393496

I. Introduction

OVER the past few decades, wireless cellular networks have been developing fast with the introduction of smart phones, tablet computers and other new mobile devices. According to a study by Cisco [1], the number of mobile devices is predicted to exceed the population on earth by the end of 2014, and by 2018, there will be nearly 1.4 mobile devices per capita. Accompanied by more data-intensive services, the global mobile data traffic is expected to increase 11-fold between 2013 and 2018, reaching 15.9 exabytes per month. The ever increasing mobile data traffic is creating challenges for current cellular network operators (CNOs). One of the promising solutions is to deploy traffic offloading with the help of complementary networks where the traffics originally targeted for mobile users (MUs) are intentionally delivered wherever and whenever possible [2], [3]. The primary objective of traffic offloading is to support more capacity-hungry services while simultaneously preserve satisfactory Quality-of-Service (QoS) for the MUs. Small cells, WiFi networks and opportunistic communications have recently emerged as the main traffic offloading technologies [2].

In this paper, we pay our attention to the problem of energy-aware traffic offloading in wireless cellular networks. As a case study, in heterogeneous cellular networks (HCNs), traffic offloading through small cells obviously alleviates much data pressure from cellular networks. Without careful designs, on the other hand, simply offloading traffic from macro cells to small cells not only may not reduce traffic congestions, but also may increase the energy consumption across the whole network. There exist some efforts on energy-aware traffic offloading in HCNs. In [5], Saker *et al.* studied the impacts of implementing femto cells with a sleep/wakeup mechanism on energy efficiency (EE). In [6], Chiang and Liao presented a reinforcement learning (RL) based scheme to intelligently offload traffic in a stochastic macro cell. These works shed lights on addressing the relation between energy saving and traffic offloading. However, none of them took into account the coupling of interference across different cells (i.e., the macro cells and the active small cells), which greatly influences the network planning and operations. The system load of a cell is referred to as the average utilization level of radio resources in both time and frequency domains [7], and is thus dependent on the system loads of other cells. In [3], Ho *et al.* provided a theoretical model for analyzing traffic offloading in load coupled heterogeneous wireless networks. In this work, we

try to explore the energy saving aspect of traffic offloading for a load coupled HCN, which is involved with the mutual interactions of various cells with one another. Meanwhile, to protect QoS for the MUs, load balancing among the cells has to be performed along with traffic offloading.

Several measurement campaigns have shown the temporal variations in mobile data traffic [4], [8], [9]. In HCNs, macro base stations (BSs) are centrally controlled by a radio network controller (or a BS controller) and manage the small cells implemented in the network. The network controller is able to obtain complete traffic information about how many MUs are associated with different macro cells and small cells given the current network state. The network state of a stochastic HCN can be hence characterized by the number of MUs in locations associated with different cells. The network state evolves according to a discrete-time Markov decision process (DTMDP), whose statistics depends on the traffic offloading strategy [5]. Hereby, a traffic offloading strategy is defined as a sequence of actions, and each action takes the form of small-cell operations (e.g., switching on a small cell to offload the traffic demands within its respective coverage or switching off a small cell to save energy). Our task is to minimize the overall energy consumption in the network while satisfying the constraint of flow-level QoS requirement in each cell such as the transmission delay [10]. In order to quantify the energy consumption and the flow-level performance, a cost criteria and the system load of each cell are associated with each state-action pair. Such a minimization problem naturally falls within the realm of a DTMDP. However, the huge number of network states hinders the application of a model-based algorithm.

We choose a model-free RL technique to solve the optimal traffic offloading strategy for a HCN with time-varying traffic. As a main contribution of this paper, we design an online learning framework, Q-learning with compact state representation (named as QC-learning), with which the network controller and the macro BSs learn to decide where and when to offload the traffic demands from MUs for the purpose of saving energy in a plausible way. The proposed approach allows computationally efficient learning in a stochastic HCN with a large network state space that cannot be handled by model-based algorithms. The remainder of this paper is organized as follows. In next section, we review the state-of-art techniques of traffic offloading in wireless cellular networks. After the brief survey, we describe the considered system model for the case study. The energy-aware traffic offloading in a stochastic load coupled HCN is formulated as a DTMDP in Section IV. In Section V, we propose a centralized QC-learning algorithm to achieve the optimal traffic offloading strategy for the network controller. A decentralized QC-learning is further developed in Section VI for each macro BS to manage the local small-cell BSs based on the global network state information obtained from the network controller. Section VII numerically evaluates the proposed studies. Finally, Section VIII concludes this paper.

II. A Brief Survey on Traffic Offloading in Wireless Networks

This section provides a brief literature survey of main traffic offloading solutions for wireless networks, including traffic offloading through small cells, traffic offloading through WiFi networks and traffic offloading through opportunistic communications. Particularly, the energy-aware traffic offloading techniques are addressed in Section II-B.

A. Traffic Offloading Solutions

1) Traffic Offloading Through Small Cells: One of the most promising trends of emerging conventional cellular technology is the small cells [11]. Small cells are small cellular BSs which deliver wireless services to a small coverage area and are most likely to be user-installed without CNO supervision. In small-cell environments, the mobile data traffic flows over the air interface to a small-cell BS and is connected to the CNO's core network through wired backhaul connections. Compared with the macro-cell deployments, the small cells can be implemented in a much more convenient and economical way. Additionally, existing studies have shown that most of the mobile data traffic is generated indoors (homes or offices) [1], [8], [9]. The CNOs thus have the opportunities to offload heavy data MUs to small cells and provide them with seamless Quality-of-Experience. For all these reasons, the small cell is viewed as an attractive and cost-effective technology of offloading traffic from macro cells.

In literature, a huge number of research works on small cells have been carried out. However, two aspects of small-cell BSs lead to serious cross- and co-tier interference issues which greatly degrade the network performance of a HCN: (a) spectrum sharing among small cells and macro cells, and (b) unplanned installment of small-cell BSs. According to the type of access control policy, small cells can be classified into two categories: open access and closed access [12]. Small cells with closed access only provide wireless services to the pre-registered MUs. To ensure satisfactory QoS, previous works have proposed such as power control and spectrum allocation approaches to control the malicious interferences. For the uplink transmissions in a two-tier HCN, Chandrasekhar and Andrews proposed a distributed utility-based signal-to-interference-plus-noise ratio (SINR) adaptation algorithm to alleviate the cross-tier interference at the macro cell received from the co-channel femto cells [13]. A Stackelberg game was formulated to study the resource allocation in a two-tier HCN, where the macro BS protects itself by pricing the interference from femto-cell MUs [14]. Regarding the downlink transmissions, Guruacharya *et al.* modeled the power allocation problem as a Stackelberg game to maximize the capacity of each BS [15]. A decentralized spectrum allocation strategy was proposed to achieve the optimal area spectral efficiency in a two-tier HCN [16]. A macro-cell beam subset selection strategy was used in [17] to reduce the cross-tier interference in two-tier femto-cell networks.

On the contrary, open access based small cells allow arbitrary nearby MUs to access the small-cell BS. From a CNO's point of view, open access policy provides an inexpensive way to improve the capacity-density across the network. Chandrasekhar and Andrews developed in [18] an uplink capacity analysis and interference avoidance strategy for a two-tier femto-cell network. The authors showed that the proposed open access

scheme can help achieve higher network-wide area spectral efficiency which is defined as the feasible number of active femto cells and MUs per cell-site. Lu *et al.* investigated spectrum allocation in an open access femto-cell network, and the proposed algorithms achieved significant performance improvements over the previous works [19]. In order to protect the QoS for neighboring macro-cell MUs in the dead zone, Li *et al.* developed a cognitive radio enhanced resource allocation method which was shown to provide transmission rate gains for both existing femto-cell and macro-cell MUs [20]. The problem of motivating small cells to admit MUs which are originally associated with macro cells has been recently studied in [21]–[23].

In practice, the number and the locations of small cells are generally unknown, which results in unpredictable interference patterns. For such dynamic networking environments, the small cells tend to behave autonomously. A real-time multi-agent RL algorithm that optimizes the network performance by managing the interference in femto-cell networks was addressed by Giupponi *et al.* in [24]. Bennis *et al.* presented a distributed learning scheme based on the well-known Q-learning to alleviate the femto-to-macro cell cross-tier interference in femto-cell networks [25]. Inspired by evolutionary game theory and machine learning, Nazir *et al.* proposed two intelligent mechanisms for interference mitigation to support the coexistence of macro cell and femto cells [26]. A Stackelberg learning and a combined learning algorithms were developed to solve the optimal resource allocation policy for an autonomous HCN in [27] and [28], respectively.

2) Traffic Offloading Through WiFi Networks: Compared with cellular technology, WiFi provides much higher data rates but with limited service coverage and mobility. Nowadays, more and more people access wireless services via WiFi connection. WiFi seems a natural solution to traffic offloading due to the built-in WiFi functionality of mobile devices. From the wireless service provider's (WSP's) perspective, WiFi is attractive because it allows data transmission to be shifted from expensive licensed spectrum bands to free unlicensed bands. Typically, there are two types of traffic offloading via WiFi networks: on-the-spot and delayed [4]. In on-the-spot offloading, the traffic demand is transmitted over the cellular network if the MU is not within the coverage of a WiFi access point (AP). While delayed offloading has been proposed that if the WiFi AP becomes unavailable, the unfinished traffic demand can be delayed up to some pre-chosen time deadline. The data transmission is completed using cellular networks, if no WiFi is detected before the deadline. Undoubtedly, WiFi offloading reduces traffic load in cellular networks [29], [30]. However, the drawback of WiFi based traffic offloading is that the CNOs completely lose visibility of the subscribed MUs whenever they are on the WiFi networks.

A lot of work has been done on traffic offloading through WiFi networks. Mehmeti and Spyropoulos used a queueing analytical model to evaluate the performance gains achieved by the on-the-spot traffic offloading, which was expressed as a function of WiFi availability and performance, and user mobility and traffic load [31]. The benefits can be extended if MUs are willing to delay their data traffic. The same authors further analyzed the case of delayed offloading in [32], where based on the queueing analytical model, the mean delay and offloading efficiency were derived as a function of the user's "patience" and some other environment parameters. A similar work was done by Lee *et al.* in [4].

However, these works failed to capture the coordination between CNOs and the owners of WiFi networks. Several recent works addressed the network economics of traffic offloading through WiFi networks using game theory [33]–[36]. Specifically, in [33], Gao *et al.* applied a general one-to-many bargaining framework to study the economic incentive issues in the problem of traffic offloading via third party APs. Lee *et al.* investigated economic benefits gained by delayed WiFi offloading, by modelling a market based on a two-stage sequential game between a monopoly WSP and MUs [34]. In [35], Paris *et al.* formulated the problem of traffic offloading through third party WiFi APs as a combinatorial auction and designed an innovative payment rule to preserve both individual rationality and truthfulness for those realistic scenario in which only part of the traffic can be offloaded. Zhuo *et al.* provided in [36] a reverse auction based incentive framework to motivate MUs to leverage their delay tolerance for cellular traffic offloading. Kang *et al.* investigated the problem of traffic offloading through WiFi network from a CNO's perspective, and derived corresponding traffic offloading schemes [37].

Remark 1: Small cells and WiFi networks both are viable traffic offloading solutions, we have briefly summarized their advantages and disadvantages. CNOs are able to have much larger free band for arbitrary WiFi deployments, since WiFi networks operate over unlicensed spectrum bands. On the other hand, implementing small cells requires careful planning as they operate in expensive, licensed and limited spectrum bands. But CNOs capture complete visibility of traffic flows through small cells, which is usually impossible if using WiFi networks for traffic offloading.

3) Traffic Offloading Through Opportunistic Communications: Opportunistic communications have been lately considered as an important way for offloading mobile data traffic [38]. The data to be delivered in wireless cellular networks may come from content WSPs, such as sport news, weather forecasts, movie trailers, and so on. The WSPs can benefit from the delay-tolerant nature of the non-real time applications and may deliver the data to only a small group of selected MUs, i.e., the target MUs. The target MUs then further propagate the data to other subscribed MUs if the mobile devices are within the proximity and can communicate opportunistically using WiFi or Bluetooth technology.

Device-to-device (D2D) communication which utilizes licensed band can also be employed for facilitating opportunistic communications [39], [40]. The majority of advantage of such a traffic offloading approach is that there is very little or no monetary cost associated with opportunistic communications. In [38], Han *et al.* exploited opportunistic communications to enable traffic offloading in the mobile social networks. As a special case, the authors studied the target-set selection problem for data delivery. The similar topic was addressed in D2D scenarios, where Zhang *et al.* proposed a novel approach to improve the performance of D2D communications underlaid over a cellular system, through exploring the social ties

and influence among individuals [41]. Al-Kanj *et al.* investigated the problem of offloading traffic in cellular networks by reducing the required number of long-distance channels to distribute common content to a group of MUs [42]. The optimal solution was achieved by forming D2D communication networks in which the BS sends different chunks of content to some MUs that in turn, multicast to other local MUs.

However, traffic offloading through opportunistic communications is challenging due to the factors, such as the heterogeneity of data content from WSPs, the varied demands and preferences for content from MUs, and the incentives for target MUs. In [43], Li *et al.* preliminarily established a theoretical framework to study the problem of multiple-type mobile data offloading, taking into account the heterogeneity of data content, MUs' preference and several realistic concerns from the target MUs.

B. Energy Awareness in Traffic Offloading

Basically, we may evaluate the performance of traffic offloading from either the CNOs' or the MUs' perspectives, and the ultimate goal is to achieve benefits for both. Most of the related works aforementioned have investigated the traffic offloading efficiency, that is, offloading as much traffic as possible is of high priority. How to improve the EE and how to satisfy the ever increasing appetite for mobile data services are currently two critical challenges faced by the CNOs [44], [45]. EE is always an important issue in wireless communication networks. It was shown in [4] that a considerable amount of power can be saved through WiFi offloading without using any delayed transmissions. The reason is that WiFi networks can provide a higher data rate than cellular networks, and need a shorter transmission time for the required traffic demands and thus lower power consumption. In the context of delayed offloading, Nicholson and Noble studied energy saving with the assumption that MUs can tolerate a delay for their data traffic [46]. For HCNs, the problem of energy-efficient spectrum sharing and power allocation in cognitive radio femto cells was studied in [47], where a three-stage Stackelberg game model was formulated to improve the EE. In [48], Ashraf *et al.* proposed a novel energy saving procedure for the femto BS to decide when to switch on/off.

As most mobile devices are battery-powered with limited energy capacity, several works on traffic offloading with the aim of improving EE for the MUs can be found in [29], [49], [50]. A prediction-based traffic offloading protocol was presented in [29] for offloading large and socially recommended contents from cellular networks to save energy of MUs. The CNO creates user mobility profile (UMP) for its subscribers and deploys WiFi networks in the locations that are most visited. The set of most visited locations along with the UMP is sent to the MUs so that they can predict WiFi availability. In such a way, significant amount of energy can be saved. Ra *et al.* made an effort in [49] to reduce the energy consumption in the scenario of transferring large volume of data from the phone to the infrastructure. In order to minimize the energy consumption as well as keep the average queue length finite, a stable and adaptive link selection algorithm was proposed to decide whether and when to defer a transmission. Through real-world traces and experiments on a smartphone, the derived algorithm can save 10–40% of battery capacity for some workloads. To improve the EE for MUs, a traffic offloading algorithm based on the metropolitan advanced delivery network architecture was proposed in [50]. The data traffic is offloaded to a WiFi AP as long as transmitting the same volume of data consumes less energy in the WiFi transmission than using the cellular network.

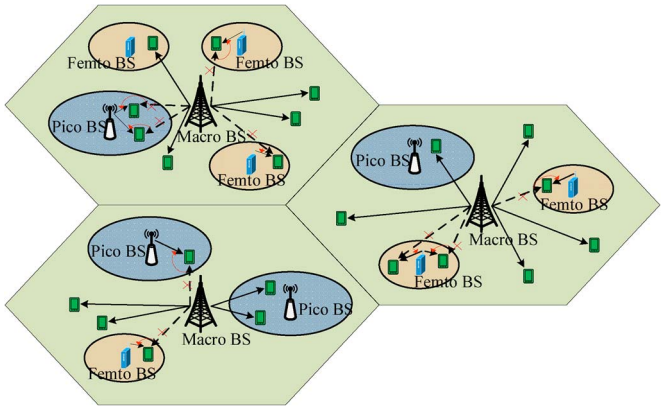

Fig. 1. An illustration example of traffic offloading in a 3-tier HCN. A macro BS can serve a MU directly or offload a MU's traffic to a nearby small cell.

III. System Description

A. Network Model

As shown in Fig. 1, this paper addresses downlink communication scenarios in a spectrum sharing HCN with multiple tiers of BSs, where each tier models a typical type of BSs. The HCN under consideration operates over discrete time epochs, each with constant time duration. The service region is represented by a set $\mathcal{L}$ of locations or small areas, each being characterized by uniform signal propagation conditions [3], [7]. At each location $l \in \mathcal{L}$ in each epoch t $(t = 1, 2, \cdots)$, the service requests follow a Poisson arrival process with arrival rate $\lambda(l, t)$. The size of requested traffic demand is assumed to be an exponentially distributed variable with mean $1/\mu(l, t)$ (in bits) [51]. Therefore, the network is Markovian [5]. A set $\mathcal{J} \triangleq \{1, \cdots, J\}$ of macro BSs ensure complete coverage, while within the coverage area of each macro BS $j \in \mathcal{J}$, K_j small-cell BSs[1] are implemented by the same CNO and are connected to the macro BS via a logical interface. $\mathcal{K}_j$ is used to denote the set of small-cell BSs in macro cell j and $\mathcal{K} \triangleq \cup_{j\in\mathcal{J}}\mathcal{K}_j$. In addition, we choose $\mathcal{L}_j^{(\mathrm{m})}$ and $\mathcal{L}_k^{(\mathrm{s})}$ to designate the sets of locations covered by macro BS j and a small-cell BS $k \in \mathcal{K}$, respectively. It is obvious that $\mathcal{L}_k^{(\mathrm{s})} \subset \mathcal{L}_j^{(\mathrm{m})}$, if $k \in \mathcal{K}_j$.

The small cells are configured to be in open access policy for the purpose of offloading traffic from the macro cells. If a small-cell BS in a macro cell is activated, the MUs appearing within the small-cell can thus sense the presence of both a macro BS and a small-cell BS, and are associated with the one that offers satisfactory QoS. The MUs connected with the macro BS will

[1]Throughout this paper, we would refer small-cell BSs to BSs of other tiers, rather than the macro BSs.

experience new source of interference but share radio resources with less users, and the MUs in other macro cells will be interfered by the active small-cell BS as well. Under this context, we aim to investigate in this paper the energy-aware traffic offloading strategy. When a macro cell is with light system load, the macro BS can serve the associated MUs alone and the small-cell BSs are switched off to save energy. For a heavily loaded macro cell, some of the small-cell BSs need to be switched on for traffic offloading [5], and the macro BS handles traffic demands from the remaining MUs that are not offloaded [52]. The working mode of a small-cell BS in each macro cell is controlled by the macro BS in a locally centralized way, based on the information of system loads in different cells. In the following discussions, we shall use interchangeably a BS and a cell.

B. Load Coupling

Traffic offloading is taken place in a particular network state at the beginning of every epoch. Let $\mathbf{x}(t) \in \mathcal{X}$ be a controlled stochastic process describing the evolution of the network state across time epochs $t = 1, 2, \cdots$. Generally, the $\mathbf{x}(t)$ can be extracted from the number of MUs of different BS at different location. To ease understanding, we rearrange the locations of all MUs served in the network. Every BS in each macro cell $j \in \mathcal{J}$ is labeled with j_k, where $k = 0, 1, \cdots, K_j$ (0 for the macro BS). And the locations covered by BS j_k are numbered from 1 to L_{j_k}. We then choose a state descriptor for $\mathbf{x}(t)$,

$$\mathbf{x}(t) = \left(\mathbf{x}_1(t), \cdots, \mathbf{x}_J(t)\right). \tag{1}$$

In (1), the $\mathbf{x}_j(t)$ describes the configuration of a local macro cell j, and is given by

$$\mathbf{x}_j(t) = \Bigg(\underbrace{x_{j_0}^1(t), \cdots, x_{j_0}^{L_{j_0}}(t)}_{\text{macro BS } j_0}, \underbrace{x_{j_1}^1(t), \cdots, x_{j_1}^{L_{j_1}}(t)}_{\text{small-cell BS } j_1}, \cdots, \underbrace{x_{j_{K_j}}^1(t), \cdots, x_{j_{K_j}}^{L_{j_{K_j}}}(t)}_{\text{small-cell BS} j_{K_j}} \Bigg), \tag{2}$$

where each element means the number of associated MUs at a location covered by a BS in macro cell j during epoch t. The subscript j_k is the label of the BS, and the superscript l ($1 \leq l \leq L_{j_k}$) is the index of its serving location. By knowing $\mathbf{x}(t)$, a traffic offloading action is selected for epoch $t+1$. An action performed in each epoch t is defined as

$$\mathbf{y}(t) = \left(\mathbf{y}_1(t), \cdots, \mathbf{y}_J(t)\right), \tag{3}$$

where each $\mathbf{y}_j(t) = (y_{j_1}(t), \cdots, y_{j_{K_j}}(t)) \in \mathcal{Y}_j$ represents the working modes of small-cell BSs within macro cell j, with $y_{j_k}(t) = 1$ if small-cell BS j_k is switched on and $y_{j_k}(t) = 0$ otherwise, for all $k \in \{1, \cdots, K_j\}$. We denote the action space by $\mathcal{Y} = \prod_{j \in \mathcal{J}} \mathcal{Y}_j$, then $\mathbf{y}(t) \in \mathcal{Y}$.

Given the network state $\mathbf{x} = \mathbf{x}(t)$ and the traffic offloading action $\mathbf{y} = \mathbf{y}(t)$ in epoch t, transmissions are scheduled for delivering the traffic demands to arriving MUs. Let $d_j^{(\mathrm{m})}(\mathbf{x}, \mathbf{y})$ and $d_k^{(\mathrm{s})}(\mathbf{x}, \mathbf{y})$ be the levels of resource utilization for macro BS $j \in \mathcal{J}$ and small-cell BS $k \in \mathcal{K}$. The average SINR achieved by MUs located at $l \in \mathcal{L}_j^{(\mathrm{m})}$ associated with macro BS j is modeled by [7]

$$\gamma_{jl}^{(\mathrm{m})}(\mathbf{x}, \mathbf{y}) = \frac{h_{jl}^{(\mathrm{m})} P_{\mathrm{tx}}^{(\mathrm{m})}}{\sum_{i \in \mathcal{J} \setminus \{j\}} h_{il}^{(\mathrm{m})} P_{\mathrm{tx}}^{(\mathrm{m})} d_i^{(\mathrm{m})}(\mathbf{x}, \mathbf{y}) + I_{jl}^{(\mathrm{m})}}, \tag{4}$$

where $I_{jl}^{(\mathrm{m})} = \sum_{k \in \mathcal{K}} h_{kl}^{(\mathrm{s})} P_{\mathrm{tx},k}^{(\mathrm{s})} d_k^{(\mathrm{s})}(\mathbf{x}, \mathbf{y}) + \delta^2$ denotes the total interference from small-cell BSs plus the background noise power, $h_{jl}^{(\mathrm{m})}$ and $h_{kl}^{(\mathrm{s})}$ are the average channel gains from macro BS j and small-cell BS k to location l, and $P_{\mathrm{tx}}^{(\mathrm{m})}$ and $P_{\mathrm{tx},k}^{(\mathrm{s})}$ are the transmit powers of a macro BS and small-cell BS k. Similarly, the received average SINR of MUs at location $l \in \mathcal{L}_k^{(\mathrm{s})}$, which is associated with an active small-cell BS k, can be expressed as

$$\gamma_{kl}^{(\mathrm{s})}(\mathbf{x}, \mathbf{y}) = \frac{h_{kl}^{(\mathrm{s})} P_{\mathrm{tx},k}^{(\mathrm{s})}}{\sum_{i \in \mathcal{K} \setminus \{k\}} h_{il}^{(\mathrm{s})} P_{\mathrm{tx},i}^{(\mathrm{s})} d_i^{(\mathrm{s})}(\mathbf{x}, \mathbf{y}) + I_{kl}^{(\mathrm{s})}}, \tag{5}$$

where $I_{kl}^{(\mathrm{s})} = \sum_{j \in \mathcal{J}} h_{jl}^{(\mathrm{m})} P_{\mathrm{tx}}^{(\mathrm{m})} d_j^{(\mathrm{m})}(\mathbf{x}, \mathbf{y}) + \delta^2$. The achievable data rates for these MUs are written as

$$R_{jl}^{(\mathrm{m})}(\mathbf{x}, \mathbf{y}) = B \log_2 \left(1 + \gamma_{jl}^{(\mathrm{m})}(\mathbf{x}, \mathbf{y})\right), \tag{6}$$

$$R_{kl}^{(\mathrm{s})}(\mathbf{x}, \mathbf{y}) = B \log_2 \left(1 + \gamma_{kl}^{(\mathrm{s})}(\mathbf{x}, \mathbf{y})\right), \tag{7}$$

in bits/second, where B is the system frequency bandwidth.

To serve the aggregated traffic demand $\vartheta(l, t)$ from MUs at location l in epoch t, the macro BS $j \in \mathcal{J}$ or the small-cell BS $k \in \mathcal{K}$ thus needs a total of $\tau_{jl}^{(\mathrm{m})}(\mathbf{x}, \mathbf{y}) \triangleq \vartheta(l, t) / R_{jl}^{(\mathrm{m})}(\mathbf{x}, \mathbf{y})$ or $\tau_{kl}^{(\mathrm{s})}(\mathbf{x}, \mathbf{y}) \triangleq \vartheta(l, t) / R_{kl}^{(\mathrm{s})}(\mathbf{x}, \mathbf{y})$ seconds. Providing that C seconds constitute one epoch in question, we obtain the system loads in macro cell j and small-cell k during epoch t by putting together the previously derived equations,

$$d_j^{(\mathrm{m})}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{l \in \bar{\mathcal{L}}_j^{(\mathrm{m})}(t)} \tau_{jl}^{(\mathrm{m})}(\mathbf{x}, \mathbf{y})}{C}, \tag{8}$$

$$d_k^{(\mathrm{s})}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{l \in \bar{\mathcal{L}}_k^{(\mathrm{s})}(t)} \tau_{kl}^{(\mathrm{s})}(\mathbf{x}, \mathbf{y})}{C}, \tag{9}$$

where $\bar{\mathcal{L}}_j^{(\mathrm{m})}(t) = \mathcal{L}_j^{(\mathrm{m})} \setminus \cup_{k \in \mathcal{K}_j} \bar{\mathcal{L}}_k^{(\mathrm{s})}(t)$ and $\bar{\mathcal{L}}_k^{(\mathrm{s})}(t)$ are the sets of locations that are associated with macro BS j and small-cell BS k during the epoch. Note that for an active small-cell BS $k \in \mathcal{K}_j$, $\bar{\mathcal{L}}_k^{(\mathrm{s})}(t) = \mathcal{L}_k^{(\mathrm{s})} \subset \mathcal{L}_j^{(\mathrm{m})}$. And $\bar{\mathcal{L}}_k^{(\mathrm{s})}(t) = \emptyset$ if small-cell BS k is switched-off, i.e., an inactive small-cell BS does not undertake any system load. From both (8) and (9), it is clear that the system load of a cell is a function of the load levels in other cells. The system load can be interpreted as the fraction of time scheduled for serving the requested traffic demands or the probability of causing interference to on-going transmissions in other cells.

IV. Problem Formulation

This section formulates the traffic offloading problem to be discussed. The MUs arrive in the network over epochs and generate service requests, and the BSs serve the requested traffic demands subject to the network conditions. A heavily loaded macro cell provides poor QoS for the associated MUs. In this case, some of the deployed small-cell BSs need to be activated. The traffic demands within the coverage of active small cells are then offloaded to the small-cell BSs. Meanwhile, new interference will be caused by the activated small-cell BSs due to spectrum sharing. At later time, the BSs finish delivering traffic demands and the MUs depart from the network. Accordingly, the energy-aware traffic offloading operation in HCNs can be described by the MU arrival and departure processes, system load coupling among different cells, the network energy consumption and the QoS requirements from MUs.

The energy consumption over a HCN depends on the system loads in different cells. In each epoch t, if a traffic offloading action $\mathbf{y} = \mathbf{y}(t)$ is executed in a network state $\mathbf{x} = \mathbf{x}(t)$, the average energy consumption of a BS during the epoch is given by [53]

$$\begin{aligned} e_j^{(\mathrm{m})}(\mathbf{x},\mathbf{y}) &= \frac{1}{C}\left(CP_{\mathrm{cst}}^{(\mathrm{m})} + \alpha^{(\mathrm{m})} d_j^{(\mathrm{m})}(\mathbf{x},\mathbf{y}) CP_{\mathrm{tx}}^{(\mathrm{m})}\right) \\ &= P_{\mathrm{cst}}^{(\mathrm{m})} + \alpha^{(\mathrm{m})} d_j^{(\mathrm{m})}(\mathbf{x},\mathbf{y}) P_{\mathrm{tx}}^{(\mathrm{m})}, \end{aligned} \tag{10}$$

for each macro BS $j \in \mathcal{J}$, where $P_{\mathrm{cst}}^{(\mathrm{m})}$ is the constant power consumption due to the signal processing unit, and $\alpha^{(\mathrm{m})}$ is a linear transmission power dependence factor. An inactive small-cell BS with no system load consumes no power, thus the energy consumption model in (10) is changed to be

$$e_k^{(\mathrm{s})}(\mathbf{x},\mathbf{y}) = \left(P_{\mathrm{cst},k}^{(\mathrm{s})} + \alpha_k^{(\mathrm{s})} d_k^{(\mathrm{s})}(\mathbf{x},\mathbf{y}) P_{\mathrm{tx},k}^{(\mathrm{s})}\right) \mathbb{1}_{\{y_k=1\}}, \tag{11}$$

for each small-cell BS $k \in \mathcal{K}$, where $P_{\mathrm{cst},k}^{(\mathrm{s})}$ and $\alpha_k^{(\mathrm{s})}$ are the constant power consumption and the transmission power dependence factor of small-cell BS k, and $\mathbb{1}_{\{\Theta\}}$ is an indicator function that equals 1 if condition Θ is met and 0, otherwise. We cast the total network energy consumption in epoch t as

$$e(\mathbf{x},\mathbf{y}) = \sum_{j\in\mathcal{J}} \left(e_j^{(\mathrm{m})}(\mathbf{x},\mathbf{y}) + \sum_{k\in\mathcal{K}_j} e_k^{(\mathrm{s})}(\mathbf{x},\mathbf{y}) \right), \tag{12}$$

which is accumulated over all macro cells.

One important metric for measuring QoS is the local system load $d_u^{(\upsilon)}(\mathbf{x}(t),\mathbf{y}(t))$, where $u \in \mathcal{J}\cup\mathcal{K}$, $\upsilon \in \{\mathrm{m},\mathrm{s}\}$ and $t = 1, 2, \cdots$. Based on the analysis in Section III-B, the value of $d_u^{(\upsilon)}(\mathbf{x}(t),\mathbf{y}(t))$ grows with the traffic demand and the amount of inter-cell interference from other active cells. Intuitively, a low system load suggests that the BS works with large sparse throughput and is able to offer the associated MUs with good capacity to meet the traffic demands, whilst a high system load indicates poor QoS in terms of congestion and potential service outage. For the latter case, the network controller should revise the network operation, through activating or deactivating small-cell BSs. Our overarching goal is to find a traffic offloading strategy $\omega : \mathcal{X} \to \mathcal{Y}$ that chooses the correct working modes $\mathbf{y}(t)$ for all small-cell BSs in every network state $\mathbf{x}(t)$, such that the energy consumption of the whole network is minimized subject to the QoS constraints. Formally, we consider the following optimization problem of finding a traffic offloading strategy ω that

$$\min_{\omega\in\Omega} \quad V(\omega) \tag{13a}$$

$$\text{s.t.} \quad d_u^{(\upsilon)}\left(\mathbf{x}(t),\mathbf{y}(t)\right) \le d_u^{\mathrm{th}},\ \forall u,\upsilon,t. \tag{13b}$$

Here Ω is the set of all available traffic offloading strategies. $V(\omega)$ defines the long-term expected energy consumption of the network under strategy ω. A predefined threshold d_u^{th} is introduced for each cell's system load in each epoch, the incentive of which is to incorporate the flow-level performance when transmission delay is a concern [10]. For example, with a small threshold value, the BS operates with a low level of radio resource consumption on average. As a result, the associated MUs would experience better throughput and thus less delay. On the other hand, a large threshold value might achieve more energy saving but leads to QoS reduction for the MUs [51].

In order to solve the optimal traffic offloading strategy, we define a DTMDP that associates to every network state an action, a corresponding state transition and a cost function. The state transitions and actions are taken place at discrete time epochs. The network controller observes in current epoch t the network state $\mathbf{x}(t)$ associated an action $\mathbf{y}(t)$, which is defined as the traffic offloading operation. The action $\mathbf{y}(t)$ is chosen from previous state and is performed on the arrival to current state. A feedback of cost is generated for the network controller in the end of the epoch. The cost function is selected to be the $e(\mathbf{x}(t),\mathbf{y}(t))$, as defined in (12). The formal expression for the DTMDP is given by $\langle \mathcal{X}, \mathcal{Y}, T, e\rangle$, where $T : \mathcal{X}\times\mathcal{Y}\times\mathcal{X} \to [0,1]$ is a state transition probability function. However, with the state descriptor as in (1), the size of $\mathcal{X}$ creates dramatic implementation difficulty. From the definitions given in (8) and (9), the system load in a cell depends on the network state. Inversely, the state space $\mathcal{X}$ can be defined by the QoS constraint in (13b) as,

$$\mathcal{X} = \{\mathbf{x} :\ \text{Constraint } (13b) \text{ holds.}\}. \tag{14}$$

Further, we may notice that the actions available during an epoch also depend on the network state. For example, if there are no MUs coming in all small cells, then all the small-cell BSs should be switched off. By eliminating the states with only one available action, we have a new reduced state space $\tilde{\mathcal{X}}$. The DTMDP is updated to $\langle \tilde{\mathcal{X}}, \mathcal{Y}, T, e\rangle$.

Ultimately, the objective of a DTMDP learner is then to find an optimal traffic offloading action $\omega^*(\mathbf{x}) \in \mathcal{Y}$ for each $\mathbf{x} \in \tilde{\mathcal{X}}$, such that a certain cumulative measure of costs $e(\mathbf{x}(t),\mathbf{y}(t))$ received over time epochs is minimized. A particular measure, which is referred to as the total expected discounted energy consumption over an infinite time horizon conditioned on initial network state $\mathbf{x}(1)$, is given by

$$V_\omega(\mathbf{x}(1)) = \mathbb{E}_\omega\left\{ \sum_{t=1}^{\infty} \beta^{t-1} e\left(\mathbf{x}(t),\omega(\mathbf{x}(t))\right) \middle| \mathbf{x}(1) \right\}, \tag{15}$$

for a traffic offloading strategy $\omega \in \Omega$, where the expectation $\mathbb{E}_\omega$ is over different actions in different network states for $t = 1, 2, \cdots$, and $\beta \in [0, 1)$ is the discount factor. (15) is generally called the value function of state $\mathbf{x}(1)$.

V. Learning Strategy With Coordination

From this section, we proceed to discuss how to come up with an optimal traffic offloading strategy in order to minimize the long-term energy consumption over the network. We first suppose that the traffic offloading is performed by the network controller in a centralized way.

A. Reinforcement Learning

When the network is in state $\mathbf{x}(t) \in \tilde{\mathcal{X}}$ in epoch t, a finite number of possible actions which are elements of the action space $\mathcal{Y}$ can be selected to perform. Let $\mathbf{y}(t)$ be the action chosen by the network controller in epoch t. For a given traffic offloading strategy $\omega \in \Omega$, the evolution of the DTMDP is Markovian with state transition probability

$$T(\mathbf{x}, \mathbf{y}, \mathbf{x}') = \Pr\left(\mathbf{x}(t+1) = \mathbf{x}' | \mathbf{x}(t) = \mathbf{x}, \mathbf{y}(t) = \mathbf{y}\right), \quad (16)$$

for $\mathbf{x}' \in \tilde{\mathcal{X}}$ and $t = 1, 2, \cdots$. Theoretically, the state transition probabilities, $T(\mathbf{x}, \mathbf{y}, \mathbf{x}')$, can be derived in a similar way as in [5], which depends on the arrival and departure rates of MUs. But an exact model of $T(\mathbf{x}, \mathbf{y}, \mathbf{x}')$ is often infeasible for the considered traffic offloading problem due to three technical reasons. First, the arrival rate of service requests from MUs, $\lambda(l, t)$, is not only location dependent, but also constrained by the traffic load situation of the network (as indicated in (13b)). Second, even for a HCN with reasonable area size, it is impossible to explicitly list all the $(\mathbf{x}, \mathbf{y}, \mathbf{x}')$ pairs. Finally, it is not always a good option to predefine a state transition model for the problem solving, the actual traffic condition deviates from the model due to the bursty nature of MU behaviors. For all these reasons, it would be really challenging to determine an exact state transition model for a practical DTMDP with large state space to compute the optimal traffic offloading strategy through applying a model-based dynamic programming algorithm. This motivates us to study in this paper a model-free solution, such as RL. RL is based on the principle that an intelligent agent tries different actions in different states infinite number of times so it can gradually adapt to the dynamically changing environment according to the received feedbacks.

Among different kinds of RL implementations, we choose the commonly used Q-learning [54] in this work. The network controller learns an optimal traffic offloading strategy by adopting the Q-learning algorithm. That is, given optimal Q-values, $Q^*(\mathbf{x}, \mathbf{y})$, the strategy ω^* defined by

$$\omega^*(\mathbf{x}) = \arg\min_{\mathbf{y} \in \mathcal{Y}} Q^*(\mathbf{x}, \mathbf{y}), \quad (17)$$

is optimal. In particular, the definition in (17) implies the following procedures: when a MU arrives, the Q-value of an optimal traffic offloading action is determined. To learn $Q^*(\mathbf{x}, \mathbf{y})$, we update the Q-value function on a transition from state $\mathbf{x}$ to state $\mathbf{x}'$ under action $\mathbf{y}$ in epoch t

$$Q^{t+1}(\mathbf{x}, \mathbf{y}) = Q^t(\mathbf{x}, \mathbf{y}) + \zeta^t \delta^t, \quad (18)$$

where $\zeta^t \in (0, 1]$ is the learning rate and

$$\delta^t = e(\mathbf{x}, \mathbf{y}) + \beta \min_{\mathbf{y}' \in \mathcal{Y}} Q^t(\mathbf{x}', \mathbf{y}') - Q^t(\mathbf{x}, \mathbf{y}), \quad (19)$$

is the temporal difference in epoch t. The initial values of $Q(\mathbf{x}, \mathbf{y})$, for all $(\mathbf{x}, \mathbf{y}) \in \tilde{\mathcal{X}} \times \mathcal{Y}$, could be arbitrary. The Q-learning is a Robbins–Monro stochastic approximation method that solves the so-called Bellman's optimality equation associated with the DTMDP. It is clear that Q-learning does not require the explicit state transition probability model, $T(\mathbf{x}, \mathbf{y}, \mathbf{x}')$. For a finite DTMDP, Q-learning algorithm ensures convergence with probability (w.p.) one to the optimal solution as $t \to \infty$ if $\sum_{t=1}^{\infty} \zeta^t$ is infinite, $\sum_{t=1}^{\infty} (\zeta^t)^2$ is finite and all state-action pairs are visited infinitely often [54]. The last condition can be satisfied if the probability of choosing any action in any state is non-zero (*exploration*). Meanwhile, the controller has to exploit the current knowledge in order to perform well (*exploitation*). A classical way to balance the trade-off between *exploration* and *exploitation* is the ϵ-greedy strategy [55].

B. Learning With Compact State Representation

The Q-learning algorithm deals with the curse of modeling effectively, i.e., a state transition model is not required during evolution of the DTMDP. Another challenging issue with DTMDP is the curse of dimensionality, which means the computational complexity increases along with the sizes of the state and action spaces. In above treatment, it is supposed that the state space is small enough so that we can apply a simple lookup table, where a separate $Q(\mathbf{x}, \mathbf{y})$ is kept for each state-action pair $(\mathbf{x}, \mathbf{y})$. Obviously, when the number of state-action pairs becomes extremely large, explicitly representing each $Q(\mathbf{x}, \mathbf{y})$ becomes impossible, and a form of compact representation where the Q-values are approximated as a function of a much smaller set of variables is needed. We restrict our attention to the case where the reduced state space $\tilde{\mathcal{X}}$ is countable, and consider compact representations of $Q : \tilde{\mathcal{X}} \times \mathcal{Y} \to \mathbb{R}$ using a function $\bar{Q} : \tilde{\mathcal{X}} \times \mathcal{Y} \times \mathbb{R}^N \to \mathbb{R}$ which is referred to as a function approximator. To approximate the Q-functions, a parameter vector $\boldsymbol{\varphi} = [\{\varphi_n\}_{n=1}^N] \in \mathbb{R}^N$ is usually adopted so as to minimize a certain metric of difference between functions $Q^*(\mathbf{x}, \mathbf{y})$ and $\bar{Q}(\mathbf{x}, \mathbf{y}, \boldsymbol{\varphi})$, for all $(\mathbf{x}, \mathbf{y}) \in \tilde{\mathcal{X}} \times \mathcal{Y}$. In the special case of linear representation, the approximated function $\bar{Q}$ takes the form of

$$\bar{Q}(\mathbf{x}, \mathbf{y}, \boldsymbol{\varphi}) = \sum_{n=1}^{N} \varphi_n \phi_n(\mathbf{x}, \mathbf{y}) = \boldsymbol{\varphi} \boldsymbol{\phi}^\top(\mathbf{x}, \mathbf{y}), \quad (20)$$

where $\top$ denotes a transpose operator and the vector $\boldsymbol{\phi}(\mathbf{x}, \mathbf{y}) = [\{\phi_n(\mathbf{x}, \mathbf{y})\}_{n=1}^N]$ with each $\phi_n(\mathbf{x}, \mathbf{y})$ denoting a fixed scalar function defined over $\tilde{\mathcal{X}} \times \mathcal{Y}$. The functions $\phi_n(\mathbf{x}, \mathbf{y})$ $(n = 1, \cdots, N)$ can be viewed as the basis functions (BFs), and the φ_n $(n = 1, \cdots, N)$ as the associated weights.

The Q-learning introduced in Section V-A can be combined with linearly parameterized compact state representation via using gradient-based updates, which is named as QC-learning. The QC-learning algorithm updates the parameter vector with

$$\boldsymbol{\varphi}^{t+1} = \boldsymbol{\varphi}^t + \zeta^t \tilde{\delta}^t \nabla \bar{Q}^t(\mathbf{x}, \mathbf{y}, \boldsymbol{\varphi}^t), \tag{21}$$

over epochs, where $\boldsymbol{\varphi}^t = [\{\varphi_n^t\}_{n=1}^N]$ is the vector of parameter value in epoch t, $\tilde{\delta}^t$ is a generic temporal difference in the epoch and is defined by the following approximation of the temporal difference in traditional Q-learning (19),

$$\tilde{\delta}^t = e(\mathbf{x}, \mathbf{y}) + \beta \min_{\mathbf{y}' \in \mathcal{Y}} \boldsymbol{\varphi}^t \boldsymbol{\phi}^\top(\mathbf{x}', \mathbf{y}') - \boldsymbol{\varphi}^t \boldsymbol{\phi}^\top(\mathbf{x}, \mathbf{y}), \tag{22}$$

and the gradient is a vector of partial derivatives with respect to the elements of $\boldsymbol{\varphi}^t$ and is expressed by

$$\nabla \bar{Q}^t(\mathbf{x}, \mathbf{y}, \boldsymbol{\varphi}^t) = \boldsymbol{\phi}(\mathbf{x}, \mathbf{y}). \tag{23}$$

Notice that the updating rule in (21) is performed on a vector basis. In general, the obtained QC-learning does not converge [56]. To ensure the convergence, we will resort to an ordinary differential equation (ODE) in the following section to acquire the necessary conditions.

C. Convergence Properties

In this section, we establish the necessary conditions that ensure the convergence of QC-learning. We begin by introducing some notation that will make our discussion below more concise. We first define a matrix $\boldsymbol{\Phi}$ as

$$\boldsymbol{\Phi} = \mathbb{E}_\omega \left\{ \boldsymbol{\phi}^\top(\mathbf{x}, \mathbf{y}) \boldsymbol{\phi}(\mathbf{x}, \mathbf{y}) \right\}. \tag{24}$$

For a given parameter vector $\boldsymbol{\varphi}$ and a particular network state $\mathbf{x} \in \tilde{\mathcal{X}}$, we then define a vector $\boldsymbol{\phi}(\mathbf{x}; \boldsymbol{\varphi}) = [\{\phi_n(\mathbf{x}, \mathbf{y})\}_{n=1}^N]$, where $\mathbf{y} \in \mathcal{Y}_{\mathbf{x}}^{\boldsymbol{\varphi}}$ with

$$\mathcal{Y}_{\mathbf{x}}^{\boldsymbol{\varphi}} = \left\{ \mathbf{y} \in \mathcal{Y} \,\middle|\, \mathbf{y} = \arg\min_{\mathbf{y}' \in \mathcal{Y}} \boldsymbol{\varphi} \boldsymbol{\phi}^\top(\mathbf{x}, \mathbf{y}') \right\}, \tag{25}$$

denoting the set of optimal traffic offloading actions in $\mathbf{x}$. We now further define a $\boldsymbol{\varphi}$-dependent matrix

$$\boldsymbol{\Phi}^{\boldsymbol{\varphi}} = \mathbb{E}_\omega \left\{ \boldsymbol{\phi}^\top(\mathbf{x}; \boldsymbol{\varphi}) \boldsymbol{\phi}(\mathbf{x}; \boldsymbol{\varphi}) \right\}. \tag{26}$$

From definitions, we can find that both $\boldsymbol{\Phi}$ and $\boldsymbol{\Phi}^{\boldsymbol{\varphi}}$ are positive definite. Finally, we introduce the required assumptions.

Assumption 1: The BFs $\{\phi_n(\mathbf{x}, \mathbf{y})\}_{n=1}^N$ are linearly independent, for all $(\mathbf{x}, \mathbf{y}) \in \tilde{\mathcal{X}} \times \mathcal{Y}$.

Assumption 2: For every $n \in \{1, \cdots, N\}$, the BF $\phi_n(\mathbf{x}, \mathbf{y})$ is bounded, i.e., $\mathbb{E}\{\phi_n^2(\mathbf{x}, \mathbf{y})\} < \infty$; moreover, the cost $e(\mathbf{x}, \mathbf{y})$ satisfies $\mathbb{E}\{e^2(\mathbf{x}, \mathbf{y})\} < \infty$ as well.

Assumption 3: The learning rate sequence $\{\zeta^t\}_{t=1}^\infty$ satisfies $\sum_{t=1}^\infty \zeta^t = \infty$ and $\sum_{t=1}^\infty (\zeta^t)^2 < \infty$.

One of the main results in this paper accordingly follows.

Theorem 1: Under Assumptions 1–3, the QC-learning algorithm converges w.p. one, if

$$\boldsymbol{\Phi}^{\boldsymbol{\varphi}} < \boldsymbol{\Phi}, \text{ for all } \boldsymbol{\varphi} \in \mathbb{R}^N. \tag{27}$$

Proof: The convergence of the QC-learning algorithm can be analyzed in terms of the stability of the fixed points of the associated ODE, which can be written as [57]

$$\dot{\boldsymbol{\varphi}}^t = \mathbb{E}_\omega \left\{ \left(e(\mathbf{x}, \mathbf{y}) + \beta \boldsymbol{\varphi}^t \boldsymbol{\phi}^\top(\mathbf{x}'; \boldsymbol{\varphi}^t) - \boldsymbol{\varphi}^t \boldsymbol{\phi}^\top(\mathbf{x}, \mathbf{y}) \right) \boldsymbol{\phi}(\mathbf{x}, \mathbf{y}) \right\}. \tag{28}$$

If there exist a globally asymptotically stable point in the ODE defined by (28), the QC-learning algorithm converges w.p. one. Denote then by $\boldsymbol{\varphi}_1^t$ and $\boldsymbol{\varphi}_2^t$ two distinct trajectories of the ODE possibly starting with different initial conditions, and designate $\boldsymbol{\varphi}_0^t = \boldsymbol{\varphi}_1^t - \boldsymbol{\varphi}_2^t$, we have

$$\begin{aligned} \frac{\partial \left\| \boldsymbol{\varphi}_0^t \right\|_2^2}{\partial t} &= 2 \left(\dot{\boldsymbol{\varphi}}_1^t - \dot{\boldsymbol{\varphi}}_2^t \right) \left(\boldsymbol{\varphi}_0^t \right)^\top \\ &= 2\beta \mathbb{E}_\omega \left\{ \boldsymbol{\varphi}_1^t \boldsymbol{\phi}^\top \left(\mathbf{x}'; \boldsymbol{\varphi}_1^t \right) \boldsymbol{\phi}(\mathbf{x}, \mathbf{y}) \left(\boldsymbol{\varphi}_0^t \right)^\top \right. \\ &\quad \left. - \boldsymbol{\varphi}_2^t \boldsymbol{\phi}^\top \left(\mathbf{x}'; \boldsymbol{\varphi}_2^t \right) \boldsymbol{\phi}(\mathbf{x}, \mathbf{y}) \left(\boldsymbol{\varphi}_0^t \right)^\top \right\} - 2 \boldsymbol{\varphi}_0^t \boldsymbol{\Phi} \left(\boldsymbol{\varphi}_0^t \right)^\top. \end{aligned} \tag{29}$$

From the definition of $\boldsymbol{\phi}(\mathbf{x}; \boldsymbol{\varphi})$ given in (29), it is straightforward to state that

$$\boldsymbol{\varphi}_1^t \boldsymbol{\phi}^\top \left(\mathbf{x}'; \boldsymbol{\varphi}_1^t \right) \le \boldsymbol{\varphi}_1^t \boldsymbol{\phi}^\top \left(\mathbf{x}'; \boldsymbol{\varphi}_2^t \right), \tag{30}$$

$$\boldsymbol{\varphi}_2^t \boldsymbol{\phi}^\top \left(\mathbf{x}'; \boldsymbol{\varphi}_2^t \right) \le \boldsymbol{\varphi}_2^t \boldsymbol{\phi}^\top \left(\mathbf{x}'; \boldsymbol{\varphi}_1^t \right). \tag{31}$$

Since the expectation $\mathbb{E}_\omega$ in (29) is taken over different traffic offloading actions in different network states, we would define the sets $\mathcal{W}_+ = \{(\mathbf{x}, \mathbf{y}) \in \tilde{\mathcal{X}} \times \mathcal{Y} \mid \boldsymbol{\varphi}_0^t \boldsymbol{\phi}^\top(\mathbf{x}, \mathbf{y}) > 0\}$ and $\mathcal{W}_- = \tilde{\mathcal{X}} \times \mathcal{Y} - \mathcal{W}_+$. Combining (30) and (31), (29) can then be rewritten as

$$\begin{aligned} \frac{\partial \left\| \boldsymbol{\varphi}_0^t \right\|_2^2}{\partial t} &\le 2\beta \left(\mathbb{E}_\omega \left\{ \boldsymbol{\varphi}_0^t \boldsymbol{\phi}^\top \left(\mathbf{x}'; \boldsymbol{\varphi}_2^t \right) \boldsymbol{\phi}(\mathbf{x}, \mathbf{y}) \left(\boldsymbol{\varphi}_0^t \right)^\top \middle| \mathcal{W}_+ \right\} \right. \\ &\quad \left. + \mathbb{E}_\omega \left\{ \boldsymbol{\varphi}_0^t \boldsymbol{\phi}^\top \left(\mathbf{x}'; \boldsymbol{\varphi}_1^t \right) \boldsymbol{\phi}(\mathbf{x}, \mathbf{y}) \left(\boldsymbol{\varphi}_0^t \right)^\top \middle| \mathcal{W}_- \right\} \right) \\ &\quad - 2 \boldsymbol{\varphi}_0^t \boldsymbol{\Phi} \left(\boldsymbol{\varphi}_0^t \right)^\top. \end{aligned} \tag{32}$$

After applying Hölder's inequality [58] to each expectation on right hand side of (32), we have

$$\begin{aligned} \frac{\partial \left\| \boldsymbol{\varphi}_0^t \right\|_2^2}{\partial t} &\le 2\beta \left(\sqrt{\mathbb{E}_\omega \left\{ \left(\boldsymbol{\varphi}_0^t \boldsymbol{\phi}^\top \left(\mathbf{x}'; \boldsymbol{\varphi}_2^t \right) \right)^2 \middle| \mathcal{W}_+ \right\}} \right. \\ &\quad \times \sqrt{\mathbb{E}_\omega \left\{ \left(\boldsymbol{\phi}(\mathbf{x}, \mathbf{y}) \left(\boldsymbol{\varphi}_0^t \right)^\top \right)^2 \middle| \mathcal{W}_+ \right\}} \\ &\quad + \sqrt{\mathbb{E}_\omega \left\{ \left(\boldsymbol{\varphi}_0^t \boldsymbol{\phi}^\top \left(\mathbf{x}'; \boldsymbol{\varphi}_1^t \right) \right)^2 \middle| \mathcal{W}_- \right\}} \\ &\quad \left. \times \sqrt{\mathbb{E}_\omega \left\{ \left(\boldsymbol{\phi}(\mathbf{x}, \mathbf{y}) \left(\boldsymbol{\varphi}_0^t \right)^\top \right)^2 \middle| \mathcal{W}_- \right\}} \right) - 2 \boldsymbol{\varphi}_0^t \boldsymbol{\Phi} \left(\boldsymbol{\varphi}_0^t \right)^\top \\ &\le 2\beta \left(\sqrt{\mathbb{E}_\omega \left\{ \left(\boldsymbol{\varphi}_0^t \boldsymbol{\phi}^\top \left(\mathbf{x}'; \boldsymbol{\varphi}_2^t \right) \right)^2 \right\}} \right. \\ &\quad \left. \times \sqrt{\mathbb{E}_\omega \left\{ \left(\boldsymbol{\phi}(\mathbf{x}, \mathbf{y}) \left(\boldsymbol{\varphi}_0^t \right)^\top \right)^2 \middle| \mathcal{W}_+ \right\}} \right. \end{aligned}$$

$$
\begin{aligned}
&+\sqrt{\mathbb{E}_\omega\left\{\left(\boldsymbol{\varphi}_0^t\boldsymbol{\phi}^\top\left(\mathbf{x}';\boldsymbol{\varphi}_1^t\right)\right)^2\right\}}\\
&\times\sqrt{\mathbb{E}_\omega\left\{\left(\boldsymbol{\phi}(\mathbf{x},\mathbf{y})\left(\boldsymbol{\varphi}_0^t\right)^\top\right)^2\middle|\mathcal{W}_-\right\}}\Bigg)-2\boldsymbol{\varphi}_0^t\boldsymbol{\Phi}\left(\boldsymbol{\varphi}_0^t\right)^\top\\
&\le\ 2\beta\sqrt{\max\left\{\boldsymbol{\varphi}_0^t\boldsymbol{\Phi}^{\boldsymbol{\varphi}_1^t}\left(\boldsymbol{\varphi}_0^t\right)^\top,\boldsymbol{\varphi}_0^t\boldsymbol{\Phi}^{\boldsymbol{\varphi}_2^t}\left(\boldsymbol{\varphi}_0^t\right)^\top\right\}}\\
&\times\Bigg(\sqrt{\mathbb{E}_\omega\left\{\left(\boldsymbol{\phi}(\mathbf{x},\mathbf{y})\left(\boldsymbol{\varphi}_0^t\right)^\top\right)^2\middle|\mathcal{W}_+\right\}}\\
&+\sqrt{\mathbb{E}_\omega\left\{\left(\boldsymbol{\phi}(\mathbf{x},\mathbf{y})\left(\boldsymbol{\varphi}_0^t\right)^\top\right)^2\middle|\mathcal{W}_-\right\}}\Bigg)-2\boldsymbol{\varphi}_0^t\boldsymbol{\Phi}\left(\boldsymbol{\varphi}_0^t\right)^\top\\
&=2\beta\sqrt{\max\left\{\boldsymbol{\varphi}_0^t\boldsymbol{\Phi}^{\boldsymbol{\varphi}_1^t}\left(\boldsymbol{\varphi}_0^t\right)^\top,\boldsymbol{\varphi}_0^t\boldsymbol{\Phi}^{\boldsymbol{\varphi}_2^t}\left(\boldsymbol{\varphi}_0^t\right)^\top\right\}}\\
&\times\sqrt{\mathbb{E}_\omega\left\{\left(\boldsymbol{\phi}(\mathbf{x},\mathbf{y})\left(\boldsymbol{\varphi}_0^t\right)^\top\right)^2\right\}}-2\boldsymbol{\varphi}_0^t\boldsymbol{\Phi}\left(\boldsymbol{\varphi}_0^t\right)^\top.
\end{aligned}
\tag{33}
$$

If condition in (27) is satisfied, we have

$$
\begin{aligned}
\frac{\partial\left\|\boldsymbol{\varphi}_0^t\right\|_2^2}{\partial t}&<2\beta\boldsymbol{\varphi}_0^t\boldsymbol{\Phi}\left(\boldsymbol{\varphi}_0^t\right)^\top-2\boldsymbol{\varphi}_0^t\boldsymbol{\Phi}\left(\boldsymbol{\varphi}_0^t\right)^\top\\
&=-(2-\beta)\boldsymbol{\varphi}_0^t\boldsymbol{\Phi}\left(\boldsymbol{\varphi}_0^t\right)^\top<0,
\end{aligned}
\tag{34}
$$

which means that $\boldsymbol{\varphi}_0^t$ asymptotically converges to the origin. Thus there exists one stable point of the ODE given by (28).

Therefore, we conclude that the QC-learning converges w.p. one. This completes the proof. □

Remark 2: Theorem 1 shows that there exists a globally asymptotically stable point $\boldsymbol{\varphi}^*$ of (28), which indicates that

$$
0=\mathbb{E}_\omega\left\{\left(e(\mathbf{x},\mathbf{y})+\beta\boldsymbol{\varphi}^*\boldsymbol{\phi}^\top\left(\mathbf{x}';\boldsymbol{\varphi}^t\right)-\boldsymbol{\varphi}^*\boldsymbol{\phi}^\top(\mathbf{x},\mathbf{y})\right)\boldsymbol{\phi}(\mathbf{x},\mathbf{y})\right\}.
\tag{35}
$$

That is,

$$
\boldsymbol{\varphi}^*=\mathbb{E}_\omega\left\{\left(e(\mathbf{x},\mathbf{y})+\beta\boldsymbol{\varphi}^*\boldsymbol{\phi}^\top\left(\mathbf{x}';\boldsymbol{\varphi}^*\right)\right)\boldsymbol{\phi}(\mathbf{x},\mathbf{y})\right\}\boldsymbol{\Phi}^{-1}.
\tag{36}
$$

Then the optimal approximated Q-functions verify that

$$
\begin{aligned}
\bar{Q}\left(\mathbf{x},\mathbf{y},\boldsymbol{\varphi}^*\right)=\mathbb{E}_\omega&\left\{\left(e(\mathbf{x},\mathbf{y})+\beta\boldsymbol{\varphi}^*\boldsymbol{\phi}^\top\left(\mathbf{x}';\boldsymbol{\varphi}^*\right)\right)\boldsymbol{\phi}(\mathbf{x},\mathbf{y})\right\}\\
&\times\boldsymbol{\Phi}^{-1}\boldsymbol{\phi}(\mathbf{x},\mathbf{y}),
\end{aligned}
\tag{37}
$$

for all $(\mathbf{x},\mathbf{y})\in\tilde{\mathcal{X}}\times\mathcal{Y}$.

VI. Decentralized Multi-Agent Learning

Up to now, we have discussed the feasibility of applying the derived centralized QC-learning to energy-aware traffic offloading. Even with a compact representation of the Q-value lookup table, the number of actions grows exponentially with the number of small-cell BSs implemented in the network, potentially creating practical challenges in facilitating a totally centralized traffic offloading mechanism. From a macro cell's point of view, the operations of small-cell BSs in the coverage can be managed by the local macro BS, which provides the possibility of executing a decentralized scheme. All macro BSs learn in a cooperative way to make local decisions for controlling the small-cell BSs in the service area. Therefore, the centralized traffic offloading problem in previous section turns out to be a decentralized multi-agent learning task.

In this section, we assume that all macro BSs learn cooperatively in a team Markov game, which is defined by a tuple $\mathcal{G}=\langle\mathcal{J},\tilde{\mathcal{X}},\mathcal{Y},T,e\rangle$, with the common goal of finding a joint traffic offloading strategy $\omega\in\Omega$ so as to minimize the total expected discounted energy consumption over the network which is given by (15). The optimal Q-function, $Q^*(\mathbf{x},\mathbf{y})$ for all $(\mathbf{x},\mathbf{y})\in\tilde{\mathcal{X}}\times\mathcal{Y}$, defines the optimal joint traffic offloading strategy and captures the team Markov game structure. Under each network state $\mathbf{x}\in\tilde{\mathcal{X}}$, the macro BSs play a team stage game $\mathcal{G}_\mathbf{x}=\langle\mathcal{J},\mathcal{Y},Q^*(\mathbf{x},\cdot)\rangle$ and consider the $Q^*(\mathbf{x},\cdot)$ to be independent. It is worth mentioning that different from previous discussions, the action in the team Markov game is jointly generated by the J independent macro BSs in a distributed manner. A joint traffic offloading action $\mathbf{y}$ is optimal in network state $\mathbf{x}$, if $Q^*(\mathbf{x},\mathbf{y})\le Q^*(\mathbf{x},\mathbf{y}')$ for all $\mathbf{y}'\in\mathcal{Y}$. In the case with a quite large number of network states, it is impossible to have a particular state visited infinitely often. Instead, each macro BS learns according to (21).

Corollary 1: For the team Markov game $\mathcal{G}$, the decentralized multi-agent QC-learning algorithm converges w.p. one, if the conditions in Theorem 1 hold.

Proof: Consider the J macro BSs as a single controller that follows a stationary traffic offloading strategy $\omega\in\Omega$. Then the team Markov game $\mathcal{G}$ is essentially an DTMDP as in previous section. The rest of the proof follows the proof of Theorem 1 and is omitted for brevity. □

The decentralized multi-agent learning problem then boils down to learning to coordinate. We assume that:

Assumption 4: The offloading strategies of different macro BSs do not change significantly in similar network states;

Assumption 5: The initial network state process $\{\mathbf{x}(t)\}$ evolves following a ϕ-irreducible and Harris recurrent Markov chain [59].

The similarity between two network states $\mathbf{x}$ and $\mathbf{x}'$ $(\in\tilde{\mathcal{X}})$ can be measured in terms of Hamming distance [60], which is denoted as $D_H(\mathbf{x},\mathbf{x}')$. With *Assumption 4*, each macro BS can thus conjecture the traffic offloading strategies employed by other macro BSs for current network state through making use of the knowledge from the past. The historical knowledge up to time epoch t is then given by the σ-algebra

$$
\mathcal{F}(t)=\sigma\left(\left\{\mathbf{x}(s),\mathbf{y}(s)\right\}_{s=1}^t,\left\{e\left(\mathbf{x}(s),\mathbf{y}(s)\right)\right\}_{s=1}^{t-1}\right),
\tag{38}
$$

where the information of each experienced network state $\mathbf{x}(s)$, each performed joint action $\mathbf{y}(s)$ and network energy consumption $e(\mathbf{x}(s),\mathbf{y}(s))$ can be obtained from the network controller. In each epoch t, every macro BS checks the Hamming distance between current network state $\mathbf{x}(t)$ and state $\mathbf{x}(s)$ in $\mathcal{F}(t)$, and then obtains a sample set $\hat{\mathcal{X}}_F(\mathbf{x}(t),\mathcal{F}(t))$ which includes F different most recent observations from $\mathcal{F}(t)$ that minimize $\sum_{f=1}^F D_H(\mathbf{x}(t),\mathbf{x}(s_f))$.

Next, we set up a virtual game $\mathcal{VG}_{\mathbf{x}(t)}=\langle\mathcal{J},\mathcal{Y},E(\mathbf{x}(t),\cdot)\rangle$ for a network state $\mathbf{x}(t)$ in epoch t, where $E(\mathbf{x}(t),\mathbf{y})$ is the common payoff that all macro BSs receive after performing

a joint traffic offloading action $\mathbf{y} \in \mathcal{Y}$ and is set to be 1 if $\mathbf{y} = \arg\min_{\mathbf{y}' \in \mathcal{Y}} \bar{Q}(\mathbf{x}(t), \mathbf{y}', \boldsymbol{\varphi}^*)$ and 0, otherwise. Since the macro BSs learn in a distributed manner, we choose $\tilde{\mathcal{Y}}_j(\mathbf{x}(t))$, for each macro BS j, to denote the set of joint actions that give the payoff 1 in state $\mathbf{x}(t)$. Suppose two integers, a and S, satisfy $1 \leq a \leq F \leq S$. When $t \leq S$, all macro BSs randomly control the working modes of small-cell BSs in their coverage. From epoch $t = S+1$, each macro BS j randomly picks a records $\hat{\mathcal{Y}}_{j,a}(\hat{\mathcal{X}}_F(\mathbf{x}(t), \mathcal{F}(t)))$ from the F joint actions with respect to $\hat{\mathcal{X}}_F(\mathbf{x}(t), \mathcal{F}(t))$. Let $-j$ denote all the other macro BSs in set $\mathcal{J}$ except macro BS j. If

1) there exists a joint offloading action $\mathbf{y} = (\mathbf{y}_j, \mathbf{y}_{-j}) \in \tilde{\mathcal{Y}}_j(\mathbf{x}(t))$ such that $\mathbf{y}'_{-j} = \mathbf{y}_{-j}$, for all $\mathbf{y}' = (\mathbf{y}'_j, \mathbf{y}'_{-j}) \in \hat{\mathcal{Y}}_{j,a}(\hat{\mathcal{X}}_F(\mathbf{x}(t), \mathcal{F}(t)))$;
2) there exists at least one joint action $\mathbf{y}$ such that $\mathbf{y} \in \hat{\mathcal{Y}}_{j,a}(\hat{\mathcal{X}}_F(\mathbf{x}(t), \mathcal{F}(t))) \cap \tilde{\mathcal{Y}}_j(\mathbf{x}(t))$,

then macro BS j selects an offloading action $\mathbf{y}_j(s^*)$ where $s^* = \max_s \{s | \mathbf{y}(s) \in \hat{\mathcal{Y}}_{j,a}(\hat{\mathcal{X}}_F(\mathbf{x}(t), \mathcal{F}(t))) \cap \tilde{\mathcal{Y}}_j(\mathbf{x}(t))\}$. On the other hand, if the above 1) and 2) are not satisfied, macro BS j randomly selects an action from $\breve{\mathcal{Y}}_j(\mathbf{x}(t)) \triangleq \{\mathbf{y}_j | \mathbf{y}_j = \arg\min_{\mathbf{y}'_j} \bar{E}_j(\mathbf{x}(t), \mathbf{y}'_j)\}$, where

$$\bar{E}_j(\mathbf{x}(t), \mathbf{y}_j) = \sum_{\mathbf{y}_{-j}} E(\mathbf{x}(t), \mathbf{y}) \frac{A_j^t(\mathbf{x}(t), \mathbf{y}_{-j})}{a}, \tag{39}$$

is calculated using a records randomly drawn from the F most recently performed actions. Herein, $A_j^t(\mathbf{x}(t), \mathbf{y}_{-j})$ denotes the number of times that other macro BSs perform joint action $\mathbf{y}_{-j}$ in state $\mathbf{x}(t)$.

The basic decentralized multi-agent QC-learning for each macro BS $j \in \mathcal{J}$ is accordingly described as follows.

Algorithm 1 Decentralized QC-learning for Traffic Offloading

Initialization: set $t = 1$, $\varphi_n^t \leftarrow 0$, for all $n = 1 \cdots, N$.
Learning: given current network state $\mathbf{x}(t)$.
1. If $t < S+1$, Then
   1.1. Randomly select a traffic offloading action.
2. Else
   2.1. Update $\tilde{\mathcal{Y}}_j(\mathbf{x}(t)) = \{\mathbf{y} | E(\mathbf{x}(t), \mathbf{y}) = 1\}$ for $\mathbf{x}(t)$.
   2.2. With an *exploitation* probability $1 - \epsilon$,
      2.2.1. randomly select $\hat{\mathcal{Y}}_{j,a}(\hat{\mathcal{X}}_F(\mathbf{x}(t), \mathcal{F}(t)))$ out of F joint actions associated with $\hat{\mathcal{X}}_F(\mathbf{x}(t), \mathcal{F}(t))$;
      2.2.2. calculate $\bar{E}_j(\mathbf{x}(t), \mathbf{y}_j)$ according to (39), and construct $\breve{\mathcal{Y}}_j(\mathbf{x}(t))$;
      2.2.3. if *1)* and *2)* are met, choose the most recent action from $\hat{\mathcal{Y}}_{j,a}(\hat{\mathcal{X}}_F(\mathbf{x}(t), \mathcal{F}(t))) \cap \tilde{\mathcal{Y}}_j(\mathbf{x}(t))$; otherwise, randomly choose an action from $\breve{\mathcal{Y}}_j(\mathbf{x}(t))$.
   2.3. With an *exploration* probability ϵ, randomly select a traffic offloading action.
3. End If
4. Observe transition $\mathbf{x}(t) \rightarrow \mathbf{x}(t+1)$ and $e(\mathbf{x}(t), \mathbf{y}(t))$.
5. Update $\boldsymbol{\varphi}^t$ according to (21).
6. Set $t \leftarrow t+1$.

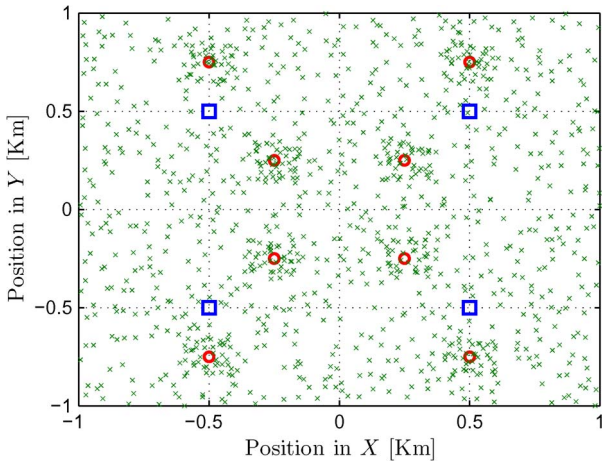

Fig. 2. A network layout: Macro BSs, small-cell BSs, MUs are, respectively, shown with blue squares, red circles and green crosses.

The following Theorem 2 ensures that the proposed decentralized multi-agent QC-learning algorithm converges to an optimal joint traffic offloading strategy.

Theorem 2: With Assumptions 1–5, the decentralized multi-agent QC-learning, described by Algorithm 1, converges w.p. one to the optimal joint traffic offloading strategy as long as for each $\mathbf{x} \in \tilde{\mathcal{X}}$, $a \leq F/(\Gamma_{\mathcal{G}_\mathbf{x}} + 2)$, where $\Gamma_{\mathcal{G}_\mathbf{x}}$ is the length of shortest path in the best response graph of team stage game $\mathcal{G}_\mathbf{x}$ [61].

Proof: The convergence of sequence $\{\boldsymbol{\varphi}^t\}$ to the optimal $\boldsymbol{\varphi}^*$ arises as an immediate consequence of Corollary 1. Therefore, the virtual games $\{\mathcal{VG}_{\mathbf{x}(t)}\}$ based on $\{\boldsymbol{\varphi}^t\}$ evolve to the virtual games $\mathcal{VG}_\mathbf{x}$ that are built upon $\boldsymbol{\varphi}^*$, for $\mathbf{x} \in \tilde{\mathcal{X}}$. On the other hand, with *Assumptions 4* and *5*, the team stage game $\mathcal{G}_{\mathbf{x}(t)}$ is reduced to a team game under network states $\hat{\mathcal{X}}_F(\mathbf{x}(t), \mathcal{F}(t))$ around $\mathbf{x}(t)$. From the result in [61, Theorem 1], the J macro BSs thus coordinate an optimal traffic offloading strategy for all $\mathbf{x}(t)$ as long as $a \leq F/(\Gamma_{\mathcal{G}_{\mathbf{x}(t)}} + 2)$. This suggests that the decentralized multi-agent QC-learning will converge to the optimal joint traffic offloading strategy w.p. one. □

VII. Numerical Results

In order to examine the performance gains from the centralized and decentralized QC-learning algorithms, numerical simulations are going to be conducted.

A. Simulation Parameters

We build up a relatively simple but representative two-tier HCN which is composed of 4 macro cells and 8 small-cells in a 2×2 Km^2 square area. The network layout is depicted in Fig. 2. The macro BSs are positioned at equal distance apart, while the small-cell BSs are at fixed locations which are assumed to be within the hotspots during simulations. Without loss of generality, only femto cells are implemented in the network for traffic offloading. Each BS is placed in the centre of a cell. The radiuses of each macro cell and each femto cell

TABLE I
PARAMETER VALUES USED IN SIMULATIONS

Parameter	Value
δ^2	4×10^{-21} W/Hz
ζ^1	0.5
β	0.9
ϵ	0.01
B	10 MHz
C	60 seconds
$P_{\text{tx}}^{(\text{m})}$, $P_{\text{tx}}^{(\text{s})}$	20 W, 0.05 W
$P_{\text{cst}}^{(\text{m})}$, $P_{\text{cst}}^{(\text{s})}$	130 W, 4.8 W
$\alpha^{(\text{m})}$, $\alpha^{(\text{s})}$	4.7, 8

are supposed to be $\sqrt{2}/2$ Km and 0.1 Km. The entire service area of the network scenario is divided into 1600 locations, i.e., each location represents a small area with a resolution of 50×50 m^2. The channel propagation in each location is considered uniform, that is, MUs in the same small area are assumed to have the same channel gains. The channel gains are fixed as $h_{ul} = z_{ul}^{-\kappa}$ for all $u \in \mathcal{J} \cup \mathcal{K}$ and $l \in \mathcal{L}$, where z_{ul} is the physical distance between BS u and the centre of location l, and κ is the path loss exponent and is set to be 4 in all simulations. The values of other parameters used in simulations are listed in Table I, where the values concerning power consumption are obtained from [62]. Each time epoch is supposed to be of 60 seconds to avoid frequent switching on/off of the femto cells. Notice that $P_{\text{tx}}^{(\text{s})}$, $P_{\text{cst}}^{(\text{s})}$ and $\alpha^{(\text{s})}$ are parameters chosen for all femto BSs. The BF vector is constructed as follows. Given the network state $\mathbf{x}(t)$ and the selected traffic offloading action $\mathbf{y}(t)$ in epoch t, we assume that the number of BFs is equal to the number of locations that are within the coverage of all femto BSs, namely, $N = |\cup_{k\in\mathcal{K}} \mathcal{L}_k^{(\text{s})}| = 128$ for the network setup. We number the locations covered by all femto cells from 1 to N. Then the value of each BF can be given by

$$\phi_n(\mathbf{x}(t), \mathbf{y}(t)) = x_{j_k}^{l_n}(t) y_{j_k}(t) \mathbb{1}_{\left\{l_n \in \mathcal{L}_k^{(\text{s})}\right\}}, \tag{40}$$

for $n \in \{1, \cdots, N\}$, where $k \in \mathcal{K}_j$ and $j \in \mathcal{J}$. That is, one BF corresponds to each location in the femto cells.

B. Performance Comparisons

We begin with demonstrating the performances that the centralized and decentralized QC-learning algorithms can achieve. In simulations, the hotspots are highly loaded with an identical arrival rate $\lambda = 6\lambda_0$, where $\lambda_0 = 0.3$ MUs/epoch is the arrival rate in areas that are not covered by the femto cells. The average file size for MUs is chosen to be a constant, $1/\mu(l,t) = 6 \times 10^6$ bits, for all locations $l \in \mathcal{L}$ and time epochs $t = 1, 2, \cdots$. And we predefine a common threshold for the system loads in all cells $d_u^{\text{th}} = 0.3$, for all $u \in \mathcal{J} \cup \mathcal{K}$. The additional parameters concerning decentralized QC-learning are given as: $a = 10$, $F = 50$ and $S = F$. Fig. 3 shows the achieved total traffic load which is the system loads over all cells during the learning process. The results are compared with two non-learning strategies:

1) *Without traffic offloading*—all femto BSs in the network are switched off and only the macro BSs serve the arriving MUs;

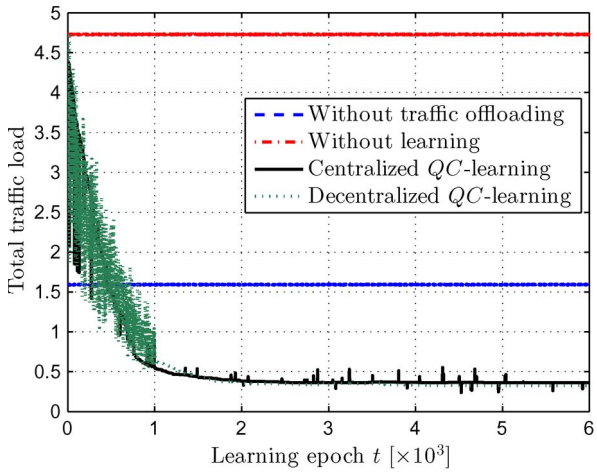

Fig. 3. Comparison of achieved system loads with respect to different traffic offloading schemes.

2) *Traffic offloading without learning*—all femto BSs are kept active all the time, such that all arriving MUs within the coverage of every femto cell are offloaded.

Without a traffic offloading implementation or offloading traffic without learning in a HCN, the evolution of the stochastic network state is given by [63]. When a learning strategy is implemented for traffic offloading, the system load in a cell depends on the network state, which reversely influences state transition probabilities. The first observation from the figure is that all curves reporting the learning processes converge within less than 3×10^3 epochs. Secondly, we may find that when all femto BSs are switched on, the achieved total system load in the network is much higher than that of when all femto BSs are switched off and those achieved by the centralized and decentralized QC-learning algorithms. This can be easily explained by the fact that activating all femto BSs for traffic offloading creates more interference in the network even when there are no MUs coming into the femto cells, thus increases traffic congestions and deteriorates the MUs' QoS. The second observation obtained from this simulation assures the necessity of designing an effective traffic offloading strategy.

To gain further insights of the proposed learning algorithms, we move on to simulate the total energy consumption over the network in each epoch during the stochastic learning process. Since keeping all femto BSs switched-on cannot ensure MUs' satisfactory QoS, we compare the energy saving performance of the centralized and decentralized QC-learning algorithms only with that of the scheme that no traffic offloading is performed. The simulation environment is the same as that used in Fig. 3. As illustrated in Fig. 4, both the centralized and decentralized QC-learning algorithms reduce total network energy consumption significantly. Another observation from the simulation results, which can also be seen from Fig. 3, is that the learning trajectories of total traffic load and total energy consumption achieved by the decentralized and centralized QC-learning algorithms are comparable. Intuitively, a locally learning macro BS obtains the global network state information from the network controller, and is thus able to asymptotically play an optimal traffic offloading action in a team Markov

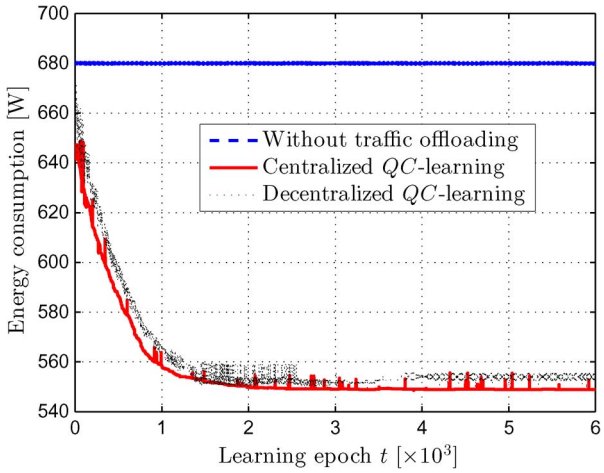

Fig. 4. Comparison of total energy consumptions in each epoch with respect to different traffic offloading schemes.

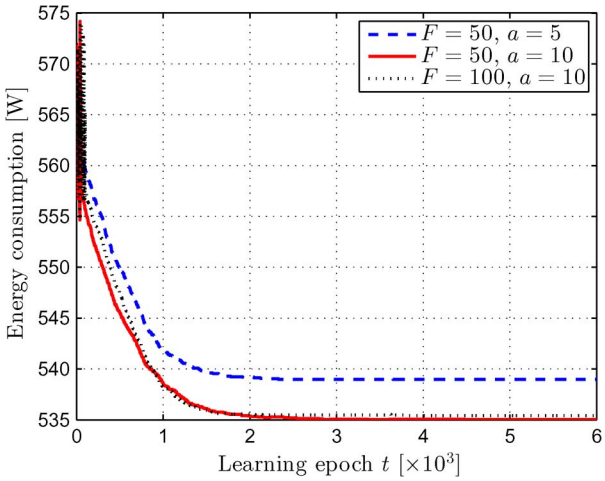

Fig. 5. Impacts of a and F on performance achieved by decentralized QC-learning.

game through making use of its historical experience. But the computational complexity of decentralized QC-learning algorithm is linear in the number of macro BSs. The exponential computational complexity of centralized QC-learning makes it infeasible for a practical scenario where the number of small-cell BSs is especially large.

The next experimental simulation studies the impacts of the parameters used in decentralized QC-learning, a and F, on the learning performance. We use similar simulation environment as in previous simulation activities except that the constant average file size for MUs is chosen to be $1/\mu(l,t) = 2 \times 10^6$ bits, for all locations $l \in \mathcal{L}$ over time epochs $t = 1, 2, \cdots$. From the plot in Fig. 5, we can identify that for a same F, worse energy saving performance is achieved if choosing a smaller value for a. The reason behind this is that each macro BS deliberately explores suboptimal traffic offloading actions during the learning process, and a smaller a increases the probability of exploring such actions. When a bigger value of F is taken, each macro BS keeps out-of-date network state information which increases the chance of exploring a suboptimal traffic offloading action.

VIII. Conclusion

In this paper, we first have presented a brief state-of-art literature review of the traffic offloading techniques that have been applied to wireless networks. Then we have focused our main emphasis on investigating a specific problem of energy-aware traffic offloading in stochastic load-coupled HCNs, the goal of which is to minimize the overall energy consumption of a network as well as to simultaneously preserve satisfactory QoS for the arriving MUs. A DTMDP was formulated to characterize the network dynamics, based on which we proposed an on-line model-free learning framework with state aggregation, i.e., the QC-learning, to solve the optimal traffic offloading strategy when the network state space is huge. Moreover, we designed a decentralized version of the centralized QC-learning algorithm for macro BSs to locally learn an optimal joint traffic offloading strategy. The convergence property of both centralized and decentralized learning algorithms was theoretically analyzed. Several experimental simulations based on a representative HCN scenario were provided in this paper to validate the proposed studies.

Acknowledgment

The authors would like to thank the anonymous reviewers for their constructive comments, which led to a significant improvement of this paper.

References

[1] "Cisco visual networking index: Global mobile data traffic forecast update, 2013–2018," San Jose, CA, USA, White Paper, Feb. 2014.
[2] A. Aijaz, H. Aghvami, and M. Amani, "A survey on mobile data offloading: Technical and business perspectives," *IEEE Wireless Commun.*, vol. 20, no. 2, pp. 104–112, Apr. 2013.
[3] C. Ho, D. Yuan, and S. Sun, "Data offloading in load coupled networks: A utility maximization framework," *IEEE Trans. Wireless Commun.*, vol. 13, no. 4, pp. 1921–1931, Apr. 2014.
[4] K. Lee, J. Lee, Y. Yi, I. Rhee, and S. Chong, "Mobile data offloading: How much can WiFi deliver?" *IEEE/ACM Trans. Netw.*, vol. 21, no. 2, pp. 536–550, Apr. 2013.
[5] L. Saker, S.-E. Elayoubi, R. Combes, and T. Chahed, "Optimal control of wake up mechanisms of femtocells in heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 664–672, Apr. 2012.
[6] Y.-H. Chiang and W. Liao, "Genie: An optimal green policy for energy saving and traffic offloading in heterogeneous cellular networks," in *Proc. IEEE ICC*, Budapest, Hungary, Jun. 2013, pp. 6230–6234.
[7] I. Siomina and D. Yuan, "Analysis of cell load coupling for LTE network planning and optimization," *IEEE Trans. Wireless Commun.*, vol. 11, no. 6, pp. 2287–2297, Jun. 2012.
[8] Y. Jin *et al.*, "Characterizing data usage patterns in a large cellular network," in *Proc. ACM CellNet*, Helsinki, Finland, Aug. 2012, pp. 7–12.
[9] M. Z. Shafiq, L. Ji, A. X. Liu, and J. Wang, "Characterizing and modeling Internet traffic dynamics of cellular networks," in *Proc. ACM SIGMETRICS*, San Joes, CA, USA, Jun. 2011, pp. 305–316.
[10] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam, "α-optimal user association and cell load balancing in wireless networks," in *Proc. IEEE INFOCOM*, San Diego, CA, USA, Mar. 2010, pp. 1–5.
[11] J. G. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. C. Reed, "Femtocells: Past, present, future," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 497–508, Apr. 2012.
[12] P. Xia, V. Chandrasekhar, and J. Andrews, "Open vs. closed access femtocells in the uplink," *IEEE Trans. Wireless Commun.*, vol. 9, no. 12, pp. 3798–3809, Dec. 2010.
[13] V. Chandrasekhar and J. G. Andrews, "Power control in two-tier femtocell networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 8, pp. 4316–4328, Aug. 2009.

[14] X. Kang, R. Zhang, and M. Motani, "Price-based resource allocation for spectrum-sharing femtocell networks: A Stackelberg game approach," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 538–549, Apr. 2012.

[15] S. Guruacharya, D. Niyato, D. I. Kim, and E. Hossain, "Hierarchical competition for downlink power allocation in OFDMA femtocell networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 4, pp. 1543–1553, Apr. 2013.

[16] V. Chandrasekhar and J. G. Andrews, "Spectrum allocation in tiered cellular networks," *IEEE Trans. Commun.*, vol. 57, no. 10, pp. 3059–3068, Oct. 2009.

[17] S. Park, W. Seo, Y. Kim, S. Lim, and D. Hong, "Beam subset selection strategy for interference reduction in two-tier femtocell networks," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3440–3449, Nov. 2010.

[18] V. Chandrasekhar and J. G. Andrews, "Uplink capacity and interference avoidance for two-tier femtocell networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 7, pp. 3498–3509, Jul. 2009.

[19] Z. Lu, T. Bansal, and P. Sinha, "Achieving user-level fairness in open-access femtocell-based architecture," *IEEE Trans. Mobile Comput.*, vol. 12, no. 10, pp. 1943–1954, Oct. 2013.

[20] L. Li, C. Xu, and M. Tao, "Resource allocation in open access OFDMA femtocell networks," *IEEE Wireless Commun. Lett.*, vol. 1, no. 6, pp. 625–628, Dec. 2012.

[21] P. Lin, J. Zhang, Q. Zhang, and M. Hamdi, "Enabling the femtocells: A cooperation framework for mobile and fixed-line operators," *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, pp. 158–167, Jan. 2013.

[22] S. Yun, Y. Yi, D. Cho, and J. Mo, "The economic effects of sharing femtocells," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 595–606, Apr. 2012.

[23] F. Pantisano, M. Bennis, W. Saad, and M. Debbah, "Spectrum leasing as an incentive towards uplink macrocell and femtocell cooperation," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 617–630, Apr. 2012.

[24] L. Giupponi, A. M. Galindo-Serrano, and M. Dohle, "From cognition to docition: The teaching radio paradigm for distributed & autonomous deployments," *Comput. Commun.*, vol. 33, no. 17, pp. 2015–2020, Nov. 2010.

[25] M. Bennis, S. Guruacharya, and D. Niyato, "Distributed learning strategies for interference mitigation in femtocell networks," in *Proc. IEEE GLOBECOM*, Houston, TX, USA, Dec. 2011, pp. 1–5.

[26] M. Nazir, M. Bennis, K. Ghaboosi, A. B. Mackenzie, and M. Latva-aho, "Learning based mechanisms for interference mitigation in self-organized femtocell networks," in *Proc. ASILOMAR*, Pacific Grove, CA, USA, Nov. 2010, pp. 1886–1890.

[27] X. Chen, H. Zhang, T. Chen, and M. Lasanen, "Improving energy efficiency in green femtocell networks: A hierarchical reinforcement learning framework," in *Proc. IEEE ICC*, Budapest, Hungary, Jun. 2013, pp. 2241–2245.

[28] X. Chen, H. Zhang, T. Chen, and J. Palicot, "Combined learning for resource allocation in autonomous heterogeneous cellular networks," in *Proc. IEEE PIMRC*, London, U.K., Sep. 2013, pp. 1061–1065.

[29] N. Ristanovic, J. Boudec, A. Chaintreau, and V. Erramilli, "Energy efficient offloading of 3G networks," in *Proc. IEEE MASS*, Valencia, Spain, Oct. 2011, pp. 202–211.

[30] S. Dimatteo, P. Hui, B. Han, and V. Li, "Cellular traffic offloading through WiFi networks," in *Proc. IEEE MASS*, Valencia, Spain, Oct. 2011, pp. 192–201.

[31] F. Mehmeti and T. Spyropoulos, "Performance analysis of "on-the-spot" mobile data offloading," in *Proc. IEEE GLOBECOM*, Atlanta, GA, USA, Dec. 2013, pp. 1577–1583.

[32] F. Mehmeti and T. Spyropoulos, "Is it worth to be patient? Analysis and optimization of delayed mobile data offloading," in *Proc. IEEE INFOCOM*, Toronto, ON, Canada, Apr./May 2014, pp. 2364–2372.

[33] L. Gao, G. Iosidis, J. Huang, L. Tassiulas, and D. Li, "Bargaining-based mobile data offloading," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1114–1125, Jun. 2014.

[34] J. Lee, Y. Yi, S. Chong, and Y. Jin, "Economics of WiFi offloading: Trading delay for cellular capacity," in *Proc. IEEE INFOCOM SDP Workshop*, Turin, Italy, Apr. 2013, pp. 357–362.

[35] S. Paris, F. Martignon, I. Filippini, and L. Chen, "A bandwidth trading marketplace for mobile data offloading," in *Proc. INFOCOM*, Turin, Italy, Apr. 2013, pp. 430–434.

[36] X. Zhuo, W. Gao, G. Cao, and S. Hua, "An incentive framework for cellular traffic offloading," *IEEE Trans. Mobile Comput.*, vol. 13, no. 3, pp. 541–555, Jan. 2013.

[37] X. Kang, Y.-K. Chia, S. Sun, and H. F. Chong, "Mobile data offloading through a third-party WiFi access point: An operator's perspective," *IEEE Trans. Wireless Commun.*, vol. 13, no. 10, pp. 5340–5351, Oct. 2014.

[38] B. Han *et al.*, "Mobile data offloading through opportunistic communications and social participation," *IEEE Trans. Mobile Comput.*, vol. 11, no. 5, pp. 821–834, May 2012.

[39] S. Andreev, A. Pyattaev, K. Johnsson, O. Galinina, and Y. Koucheryavy, "Cellular traffic offloading onto network-assisted device-to-device connections," *IEEE Commun. Mag.*, vol. 52, no. 4, pp. 20–31, Apr. 2014.

[40] A. Antonopoulos, E. Kartsakli, and C. Verikoukis, "Game theoretic D2D content dissemination in 4G cellular networks," *IEEE Commun. Mag.*, vol. 52, no. 6, pp. 125–132, Jun. 2014.

[41] Y. Zhang *et al.*, "Social network aware device-to-device communication in wireless networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 177–190, Jan. 2015.

[42] L. Al-Kanj, H. V. Poor, and Z. Dawy, "Optimal cellular offloading via device-to-device communication networks with fairness constraints," *IEEE Trans. Wireless Commun.*, vol. 13, no. 8, pp. 4628–4643, Aug. 2014.

[43] Y. Li *et al.*, "Multiple mobile data offloading through disruption tolerant networks," *IEEE Trans. Mobile Comput.*, vol. 13, no. 7, pp. 1579–1596, Jul. 2014.

[44] J. Wu, S. Rangan, and H. Zhang, *Green Communications—Theoretical Fundamentals, Algorithms, and Applications*. Boca Raton, FL, USA: CRC Press, Sep. 2012.

[45] T. Chen, Y. Yang, H. Zhang, H. Kim, and K. Horneman, "Network energy saving technologies for green wireless access networks," *IEEE Wireless Commun.*, vol. 18, no. 5, pp. 30–38, Oct. 2011.

[46] A. J. Nicholson and B. D. Noble, "BreadCrumbs: Forecasting mobile connectivity," in *Proc. ACM MobiCom*, San Francisco, CA, USA, Sep. 2008, pp. 46–57.

[47] R. Xie, F. R. Yu, and H. Ji, "Energy-efficient spectrum sharing and power allocation in cognitive radio femtocell networks," in *Proc. IEEE INFOCOM*, Orlando, FL, USA, Mar. 2012, pp. 1665–1673.

[48] I. Ashraf, L. T. W. Ho, and H. Claussen, "Improving energy efficiency of femtocell base stations via user activity detection," in *Proc. IEEE WCNC*, Sydney, Australia, Apr. 2010, pp. 1–5.

[49] M.-R. Ra *et al.*, "Energy–delay tradeoffs in smartphone applications," in *Proc. ACM MobiSys*, San Francisco, CA, USA, Jun. 2010, pp. 255–270.

[50] A. Y. Ding *et al.*, "Enabling energy-aware collaborative mobile data offloading for smartphones," in *Proc. IEEE SECON*, New Orleans, LA, USA, Jun. 2013, pp. 487–495.

[51] E. Oh, K. Son, and B. Krishnamachari, "Dynamic base station switching-on/off strategies for green cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 2126–2136, May 2013.

[52] L. Gao, G. Iosifidis, J. Huang, and L. Tassiulas, "Economics of mobile data offloading," in *Proc. IEEE INFOCOM SDP Workshop*, Turin, Italy, Apr. 2013, pp. 3303–3308.

[53] H. Holtkamp, G. Auer, S. Bazzi, and H. Haas, "Minimizing base station power consumption," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 2, pp. 297–306, Feb. 2014.

[54] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, no. 3/4, pp. 279–292, 1992.

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

[56] L. Baird, "Residual algorithms: Reinforcement learning with function approximation," in *Proc. ICML*, Tahoe City, CA, USA, Jul. 1995, pp. 30–37.

[57] M. Wunder, M. Littman, and M. Babes, "Classes of multiagent Q-learning dynamics with ϵ-greedy exploration," in *Proc. ICML*, Haifa, Israel, Jun. 2010, pp. 1167–1174.

[58] J. M. Aldaz, "A stability version of Hölder's inequality," *J. Math. Anal. Appl.*, vol. 343, no. 2, pp. 842–852, Jul. 2008.

[59] G. O. Roberts and J. S. Rosenthal, "Harris recurrence of Metropolis-within-Gibbs and trans-dimensional Markov chains," *Ann. Appl. Probab.*, vol. 16, no. 4, pp. 2123–2139, 2006.

[60] L. N. de Castro and F. J. Von Zuben, "Learning and optimization using the clonal selection principle," *IEEE Trans. Evol. Comput.*, vol. 6, no. 3, pp. 239–251, Aug. 2002.

[61] H. P. Young, "The evolution of conventions," *Econometrica*, vol. 61, no. 1, pp. 57–84, Jan. 1993.

[62] Y. S. Soh, T. Q. S. Quek, M. Kountouris, and H. Shin, "Energy efficient heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 5, pp. 840–850, May 2013.

[63] F. Baskett, K. M. Chandy, R. R. Muntz, and F. G. Palacios, "Open, closed and mixed networks of queues with different classes of customers," *J. ACM*, vol. 22, no. 2, pp. 248–260, Apr. 1975.

Xianfu Chen (M'13) received the Ph.D. degree in signal and information processing from Zhejiang University, Hangzhou, China, in 2012. In April 2012, he joined the VTT Technical Research Centre of Finland Ltd., Oulu, Finland, where he is currently a Senior Scientist. His research interests cover various aspects of wireless communications and networking, with emphasis on software-defined radio access networks, green communications, centralized and decentralized resource allocation, and the application of artificial intelligence to wireless communications.

Jinsong Wu (M'07–SM'11) is the Founder and Founding Chair of the Technical Committee on Green Communications and Computing (TCGCC), IEEE Communications Society, which was established in 2011 as an official Technical Subcommittee (TSCGCC) and elevated as TCGCC in 2013. He is an Associate Editor of IEEE COMMUNICATIONS SURVEYS AND TUTORIALS, IEEE SYSTEMS JOURNAL, and IEEE ACCESS. He is an Area Editor of the incoming IEEE JOURNAL ON SELECTED AREAS in Communications Series on Green Communications and Networking, starting in 2015. He is the Founder and Series Editor on Green Communication and Computing networks of IEEE COMMUNICATIONS MAGAZINE. He has served as a co-leading Guest Editor of Special Issue on Green Communications, Computing, and Systems in IEEE SYSTEMS JOURNAL and an Associate Editor of Special Section on Big Data for Green Communications and Computing in IEEE ACCESS. He was the leading Editor and a co-author of the comprehensive book entitled *Green Communications: Theoretical Fundamentals, Algorithms, and Applications (CRC Press).*

Yueming Cai (M'05–SM'12) received the B.S. degree in physics from Xiamen University, Xiamen, China, in 1982 and the M.S. degree in microelectronics engineering and the Ph.D. degree in communication and information systems from Southeast University, Nanjing, China, in 1988 and 1996, respectively. He is currently a Full Professor with the College of Communications Engineering, PLA University of Science and Technology, Nanjing. His research interests include cooperative communications, signal processing in communications, wireless sensor networks, and physical layer security.

Honggang Zhang (M'01–SM'11) is a Full Professor with the Department of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China. He is an Honorary Visiting Professor at the University of York, York, U.K. He was the International Chair Professor of Excellence for Université Européenne de Bretagne (UEB) and Supélec, France. He served as the Chair of the Technical Committee on Cognitive Networks of the IEEE Communications Society from 2011 to 2012. He is currently active in the research on green communications and was the leading Guest Editor of the IEEE COMMUNICATIONS MAGAZINE special issues on "Green Communications." He was the co-author and an Editor of two books with the titles of *Cognitive Communications-Distributed Artificial Intelligence (DAI), Regulatory Policy and Economics, Implementation* (John Wiley & Sons) and *Green Communications: Theoretical Fundamentals, Algorithms and Applications* (CRC Press), respectively.

Tao Chen (S'05–M'10–SM'13) received the B.E. degree in telecommunications engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 1996 and the Ph.D. degree in telecommunications engineering from the University of Trento, Trento, Italy, in 2007. He is currently a Senior Researcher at VTT Technical Research Center of Finland Ltd., Oulu, Finland. His current research interests include dynamic spectrum access, energy efficiency, and resource management in heterogeneous wireless networks, software defined networking for 5G mobile networks, and social-aware mobile networks.