# Spectrum Sharing Among Rapidly Deployable Small Cells: A Hybrid Multi-Agent Approach

Bo Gao, *Member, IEEE*, Lingyun Lu, Ke Xiong, *Member, IEEE*, Jung-Min Park, *Fellow, IEEE*, Yaling Yang, *Member, IEEE*, and Yuwei Wang

***Abstract*— On-demand deployment of small cells plays a key role in augmenting macro-cell coverage for outdoor hotspots, where user devices are brought together and intensively upload self-generated data. In this paper, we study spectrum sharing among rapidly deployable small cells in the uplink, even without a priori global knowledge. We propose a hybrid multi-agent approach, which allows a leading macro-cell base station (MBS) and multiple following small base stations (SBSs) to take part in a user-centric, online joint optimization of small cell deployment and uplink resource allocation. Specifically, we propose a centralized mechanism for the MBS to solve the first subproblem of small cell deployment stage by stage, based on an adversarial bandit model. Furthermore, we propose a distributed mechanism for the group of SBSs to collectively solve the second subproblem of uplink resource allocation stage by stage, based on a stochastic game model. We prove that our approach is guaranteed to produce a joint strategy, which is built upon a mixed strategy with bounded regret on the first tier and an equilibrium solution on the second tier. Our approach is validated by simulations on the aspects of convergence behavior, strategy correctness, power consumption, and spectral efficiency.**

***Index Terms*— Mobile communication systems, distributed networks, algorithm/protocol design and analysis.**

## I. Introduction

IN THE era of 5G mobile communications and Internet of Things (IoT), each macro-cell can be densely overlaid with small cells, so that mobile users or connected things are served by a heterogeneous network (HetNet) [1], [2]. It is straightforward to deploy small cells for indoor hotspots such as homes or offices, but usually not for outdoor hotspots to support, e.g., public gatherings or sporting events. As the use of wireless devices outdoors is becoming a daily necessity, it has received a growing interest in augmenting macro-cell coverage for target spots that are wide-open yet short-lived [3], [4]. A further improvement of outdoor coverage is necessary when user devices are brought together in a short time and cause a burst increase in macro-cell traffic, thus leading to a supply-demand mismatch. It commonly occurs that lots of user devices are temporarily gathered for broadband or IoT services. However, it can be very costly or even impossible to fully and persistently cover an outdoor hotspot with small cells.

Recent efforts have been made to find solutions towards fast, cost-effective, and on-demand deployment of small cells outdoors. A promising innovation is to mount small base stations (SBSs) on movable and controllable platforms, such as unmanned aerial vehicles (UAVs) or unmanned ground vehicles (UGVs) [5]–[8]. Under vehicular mobility, UAV/UGV-mounted small cells are rapidly deployable in case macro-cell base stations (MBSs) are overloaded, or even damaged or destroyed. Lately, research groups from Google, Facebook, Nokia, and academic institutions have made progress in prototyping UAV/UGV-mounted small cells [9]. The notion of rapidly deployable small cells has now become feasible.

In this paper, we focus on spectrum sharing among rapidly deployable small cells particularly for outdoor coverage in the uplink, which is emphasized when user devices intensively upload their self-generated data [3], [5]–[7]. A large volume of uplink traffic can be locally generated in an outdoor hotspot due to the proliferation of sensing-capable devices. For example, a typical downlink-to-uplink ratio of mobile devices is 10:1, but it can become 1:3 during mass events along with a tenfold increase in uplink traffic [3]. This is because of the uploading of pictures and videos to social media. Likewise, collecting raw data from sensor nodes or IoT devices can lead to very heavy uplink traffic [5], [6]. To handle intensive uploading of user/machine-generated data, limited wireless spectrum has to be shared and reused by small cells efficiently and effectively. Therefore, organizing a two-tier HetNet outdoors necessitates inter-cell spectrum sharing, and we mainly focus on co-tier spectrum sharing among small cells. This is different from a typical scenario in that

Manuscript received March 28, 2019; revised August 12, 2019; accepted September 25, 2019. Date of publication October 11, 2019; date of current version January 8, 2020. This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant 2017RC022, in part by the National Natural Science Foundation of China (NSFC) under Grant 61872028, Grant 61502457, and Grant 61732017, in part by the Beijing Intelligent Logistics System Collaborative Innovation Center under Grant BILSCIC-2018KF-10, in part by the National Science Foundation (NSF) under Grant 1547241, Grant 1547366, Grant 1563832, Grant 1642928, Grant 1822173, and Grant 1824494, and in part by the industry affiliates of the Broadband Wireless Access and Applications Center (BWAC). The associate editor coordinating the review of this article and approving it for publication was Z. Dawy. *(Corresponding author: Bo Gao.)*

B. Gao and K. Xiong are with the School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China (e-mail: bogao@bjtu.edu.cn; kxiong@bjtu.edu.cn).

L. Lu is with the School of Software Engineering, Beijing Jiaotong University, Beijing 100044, China (e-mail: lylu@bjtu.edu.cn).

J.-M. Park and Y. Yang are with the Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA 24061 USA (e-mail: jungmin@vt.edu; yyang8@vt.edu).

Y. Wang is with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: wangyuwei@ict.ac.cn).

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

Digital Object Identifier 10.1109/TWC.2019.2945712

network performance is valued more in the uplink than in the downlink.

However, we have to address the following challenges. First, our approach has to jointly optimize small cell deployment and uplink resource allocation. In addition to the 3D placement of movable small cells to cover an outdoor hotspot, it is also necessary to study the establishment of uplink transmissions to meet user needs. Specifically, the problem of inter-cell spectrum sharing involves decision making on four aspects, including small cell placement (SCP), user-cell association (UCA), channel resource allocation (CRA), and transmit power control (TPC). Second, our approach has to be responsive to initially unknown and varying demands of user devices for uplink transmissions. It is common in an outdoor hotspot that movable small cells can only selectively fulfill a part of user demands for a limited period of time, but the distribution of user demands is not known a priori and is even ever-changing. Thus, the joint optimization problem has to be solved in a user-centric, online manner. Third, our approach has to be applicable to a resource-constrained cellular system. Although the base stations in two tiers of a HetNet can work as decision-making entities, the user-centric, online joint optimization to be performed can be too complex for either a MBS or a SBS. Any entity can suffer from limited computing, energy, and storage resources. Besides, it is already hard to derive a real-time solution to such a complex problem even given sufficient resources.

In this paper, we overcome the above challenges and make the contributions as follows.

- We propose a hybrid multi-agent approach, which operates with two tiers of sequential decision making under user uncertainty. It allows a leading MBS and multiple following SBSs to take part in a multi-stage joint optimization of small cell deployment and uplink resource allocation.
- We propose a centralized mechanism for the MBS to solve the subproblem of small cell deployment (including SCP and UCA) stage by stage, based on an adversarial bandit model. The MBS refines its strategy and updates user knowledge through reinforcement learning.
- We propose a distributed mechanism for the group of SBSs to collectively solve the subproblem of uplink resource allocation (including CRA and TPC) stage by stage, based on a stochastic game model. Under the guidance of the MBS, all the SBSs refine their strategies through multi-agent reinforcement learning.
- We prove that our approach produces a joint strategy, which is built upon a mixed strategy with bounded regret on the first tier and an equilibrium solution on the second tier. Our approach is further validated by simulations on the aspects of convergence behavior, strategy correctness, power consumption, and spectral efficiency.

The remainder of this paper is organized as follows. Related work is discussed in section II. System model is presented in section III, where the original problem is formulated assuming centralized control. Based on problem decoupling, our hybrid multi-agent approach is outlined in section IV, and is further elaborated in sections V and VI. Performance evaluation is conducted in section VII. Finally, our conclusion is summarized in section VIII.

## II. Related Work

Ever since the advent of rapidly deployable solutions, there have been recent studies that investigate the placement of either a single or multiple UAV/UGV-mounted small cells. It is assumed that each mobile platform hovers or remains still in one spot for the optimal small cell coverage. For single-cell scenarios, on the one hand, the altitude [10]–[12] or target position [13]–[15] of a SBS is determined based on certain user demands. As for multi-cell scenarios, on the other hand, the target positions of SBSs are determined either when small cells occupy separate spectrum segments [16]–[18] or when small cells (and sometimes macro-cells) share same spectrum band so that co-tier (and cross-tier) interference is not negligible [19]–[23]. However, none of the above has paid attention to spectrum sharing among small cells in the uplink, which is particularly emphasized for uploading-intensive outdoor hotspots.

There have been other studies that focus on regulating the movement of either a single or multiple UAV/UGV-mounted small cells. It is then assumed that each mobile platform keeps moving constantly for the optimal small cell coverage along a trajectory. For single-cell scenarios, on the one hand, the moving trajectory of a SBS is determined to comply with certain user demands [24]–[28]. As for multi-cell scenarios, on the other hand, the moving trajectories of SBSs are determined either when small cells do not share same spectrum band [29] or when small cells coexist thus co-tier interference affects decision-making processes [30]–[33]. However, these literatures study spectrum sharing among small cells only in the downlink as well. Moreover, existing work has to assume perfect global knowledge to optimize spectrum sharing under the mobility of small cells, but in fact, a priori user knowledge is difficult to obtain for transient outdoor hotspots.

One latest trend is to characterize or predict user demands before carrying out on-demand deployment of UAV/UGV-mounted small cells. It is assumed in this case that user demands are ever-changing and are not known a priori. According to the estimation of user patterns through machine learning, multiple SBSs can be appropriately placed either when small cells do not share same spectrum band [34]–[36] or when small cells coexist [37]. However, none of the above can handle our particular challenges, i.e., user-centric, online joint optimization in support of spectrum sharing among small cells in the uplink.

## III. System Model

In this section, we describe the system model that underpins our hybrid multi-agent approach.

### A. Basic Assumptions

During the lifetime of an outdoor hotspot, user devices are temporarily brought together and settled down for broadband or IoT services. They are eager to upload locally generated data. We consider common wireless devices in this

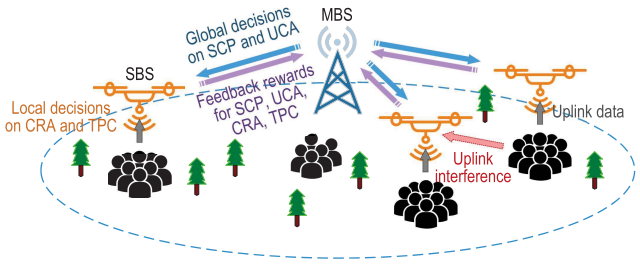


Fig. 1. Basic idea of the hybrid multi-agent approach for spectrum sharing among rapidly deployable small cells in the uplink.

scenario, e.g., event spectators' mobile devices or regular sensor nodes, whose locations and demands are slowly changing. The area of interest is covered by a two-tier HetNet, which consists of one fixed macro-cell and multiple UAV/UGV-mounted (hovering) small cells. An example of such a cellular system is illustrated in Fig. 1.

The establishment of data transmissions in the uplink relies on the allocation of certain dedicated portions of wireless spectrum. We assume that the first tier of macro-cell and the second tier of small cells operate on different spectrum bands to avoid cross-tier mutual interference. This is consistent with the regulation of 5G spectrum, in which macro-cells and small cells utilize lower and higher frequency bands, respectively [38]. Then, our primary focus can be on the tier of small cells. We assume that all the small cells operate on the same spectrum band, so that inter-cell spectrum sharing is necessary to accommodate a large volume of uplink traffic. To enable an uplink transmission, each user device can be associated with a certain home small cell and can occupy a certain fraction of shared spectrum by generating a certain level of transmit power. Isotropic antennas of user devices are assumed, which generate the worst-case co-tier mutual interference. In general, the problem of spectrum sharing involves small cell deployment and uplink resource allocation. More specifically, it involves small cell placement (SCP), user-cell association (UCA), channel resource allocation (CRA), and transmit power control (TPC).

The cellular system aims to mitigate mutual interference among rapidly deployable small cells in the uplink. We assume that user demands for uplink transmissions are not known a priori due to the temporary nature of outdoor applications. Therefore, such a situation calls for the use of sequential decision making under user uncertainty. The outdoor hotspot lasts only for a finite time horizon, which can further be divided into short stages or time periods. Then, a joint optimization problem needs to be solved sequentially in a multi-stage manner. As user knowledge is updated stage by stage, the four aspects related to spectrum sharing, i.e., SCP, UCA, CRA, TPC, should be jointly optimized. We initially assume that the above operations are fully under control of the MBS, and this assumption will be relaxed in the next section.

### *B. Problem Formulation*

Now we formulate the joint optimization problem per stage if given user knowledge. Our mathematical notation is summarized in Table I.

TABLE I
SUMMARY OF MATHEMATICAL NOTATION

| | |
|---|---|
| $\mathcal{M}$ | set of small cells |
| $\mathcal{N}$ | set of user devices |
| $\mathcal{K}$ | set of available channels |
| $T$ | time span of outdoor hotspot |
| $l_t^{(m)}$ | location point of SBS $m$ |
| $a_t^{(m,n)}$ | association of user $n$ with SBS $m$ |
| $b_t^{(m,n,k)}$ | allocation of channel $k$ to user $n$ |
| $p_t^{(m,n,k)}$ | transmit power of user $n$ on channel $k$ |
| $U_t$ | weighted sum of power consumption |
| $\lambda_t^{(n)}$ | demand of user $n$ |
| $w_t^{(m,n)}$ | weight of user $n$ |
| $H_t^{(n,m)}$ | propagation gain from user $n$ to SBS $m$ |
| $\gamma^{(n)}$ | SINR requirement of user $n$ |
| $I_t^{(m,k)}$ | interference power at SBS $m$ on channel $k$ |
| $\delta^{(n)}$ | capacity requirement of user $n$ |
| $\bar{R}_t(\cdot)$ | feedback cost for the MBS's action |
| $\pi$ | mixed strategy of the MBS |
| $\bar{G}^{\pi}/\tilde{G}^{\pi}$ | expected/pseudo- regret under strategy $\pi$ |
| $\eta$ | control factor in Algorithm 1 |
| $J$ | size of action space for UCA |
| $R_t^{(m)}(\cdot,\cdot)$ | feedback cost for SBS $m$'s action |
| $\theta^{(m)}$ | pure strategy of SBS $m$ |
| $G^{\{\theta^{(m)},\cdot\}}$ | discounted cost under strategy $\theta^{(m)}$ |
| $\varphi^{(m)}$ | price factor in Algorithm 2 |

Suppose that within the macro-cell, there are a set of small cells, say $\mathcal{M} = \{1, 2, \ldots, M\}$, to be deployed and a set of user devices, say $\mathcal{N} = \{1, 2, \ldots, N\}$, to be served. Furthermore, there are a set of available channels, say $\mathcal{K} = \{1, 2, \ldots, K\}$, to be shared by small cells. Enabling $\mathcal{M}$ to serve $\mathcal{N}$ by sharing $\mathcal{K}$ involves the issues of SCP, UCA, CRA, TPC. Hence, the MBS has to determine four sets of decision variables in every stage $t = 1, 2, \ldots, T$ for a time horizon $T$.

- First for SCP, the coordinates of a SBS are selected from a 3D location space, say $\mathcal{L} \subset \mathbb{R}^3$. To always keep up with ever-changing user needs, the MBS dynamically places each SBS $m \in \mathcal{M}$ at a certain location point $l_t^{(m)} = (x_t^{(m)}, y_t^{(m)}, z_t^{(m)}) \in \mathcal{L}$. Thus, let the first variable set be $L_t = \{l_t^{(m)} \,|\, m \in \mathcal{M}\}$.
- Second for UCA, each user can be associated with one nearby SBS as its home base station. The MBS determines a binary indicator, $a_t^{(m,n)} \in \mathcal{A} = \{0, 1\}$, for each pair of SBS $m \in \mathcal{M}$ and user $n \in \mathcal{N}$, which is equal to 1 (or 0) when user $n$ is (or is not) associated with SBS $m$. Thus, let the second variable set be $A_t = \{a_t^{(m,n)} \,|\, m \in \mathcal{M}, n \in \mathcal{N}\}$, and the set of user devices being associated with each SBS $m \in \mathcal{M}$ be $\mathcal{N}_t^{(m)} = \{n \,|\, a_t^{(m,n)} = 1 \text{ for } n \in \mathcal{N}\}$.
- Third for CRA, all the small cells share the same channel set, i.e., $\mathcal{K}$, so co-channel uplink transmissions may interfere with each other. The MBS determines another binary indicator, $b_t^{(m,n,k)} \in \mathcal{B} = \{0, 1\}$, for each combination of SBS $m \in \mathcal{M}$, user $n \in \mathcal{N}_t^{(m)}$, and channel $k \in \mathcal{K}$, which is equal to 1 (or 0) when channel $k$ is (or is not) allocated to user $n$ being associated with SBS $m$. Thus, let the third variable set be $B_t = \{b_t^{(m,n,k)} \,|\, m \in \mathcal{M}, n \in \mathcal{N}_t^{(m)}, k \in \mathcal{K}\}$ or $B_t^{(m)} = \{b_t^{(m,n,k)} \,|\, n \in \mathcal{N}_t^{(m)}, k \in \mathcal{K}\}$ for

each SBS $m \in \mathcal{M}$, and the set of available channels being allocated to each pair of SBS $m \in \mathcal{M}$ and user $n \in \mathcal{N}_t^{(m)}$ be $\mathcal{K}_t^{(m,n)} = \{k \,|\, b_t^{(m,n,k)} = 1 \text{ for } k \in \mathcal{K}\}$.

- Fourth for TPC, the transmit power of each user device is adjustable within a power space, say $\mathcal{P} \subset \mathbb{R}_+$. To support spectrum reuse in the uplink, for each combination of SBS $m \in \mathcal{M}$, user $n \in \mathcal{N}_t^{(m)}$, and channel $k \in \mathcal{K}_t^{(m,n)}$, the MBS makes user $n$ being served by SBS $m$ transmit data on channel $k$ with a certain power level $p_t^{(m,n,k)} \in \mathcal{P}$. Thus, let the fourth variable set be $P_t = \{p_t^{(m,n,k)} \,|\, m \in \mathcal{M}, n \in \mathcal{N}_t^{(m)}, k \in \mathcal{K}_t^{(m,n)}\}$ or $P_t^{(m)} = \{p_t^{(m,n,k)} \,|\, n \in \mathcal{N}_t^{(m)}, k \in \mathcal{K}_t^{(m,n)}\}$ for each SBS $m \in \mathcal{M}$.

In general, small cell deployment involves SCP and UCA, while uplink resource allocation involves CRA and TPC.

For each stage $t = 1, 2, \ldots, T$, the four sets of decision variables are determined to optimize the objective of spectrum sharing among movable small cells in the uplink. Because each user device generates not only transmit power, but also interference power in the eyes of others, it is reasonable to minimize user devices' power consumption (transmit power levels) in every small cell. The benefits are twofold: transmit power saving and mutual interference mitigation. Then, for each SBS $m \in \mathcal{M}$, the objective function to be minimized can be defined by

$$U_t^{(m)} = \sum_{n \in \mathcal{N}_t^{(m)}} \lambda_t^{(n)} w_t^{(m,n)} \sum_{k \in \mathcal{K}_t^{(m,n)}} p_t^{(m,n,k)}, \tag{1}$$

where each $\lambda_t^{(n)}$ denotes user $n$'s demand (i.e., probability) for channel access, which is assumed to be uncertain and has to be learnt stage by stage; each $w_t^{(m,n)}$ denotes user $n$'s weight (or priority) for resource allocation, if $n \in \mathcal{N}_t^{(m)}$, which characterizes the impact of the user's transmit power on spectrum sharing. To estimate user demand $\lambda_t^{(n)}$, or user $n$'s probability of uplink transmissions in stage $t$, traffic patterns of user $n$ over past stages can be utilized to compute user $n$'s frequency of uplink transmissions by stage $t$. To define user weight $w_t^{(m,n)}$, interference potentials of user $n$ can be evaluated. Some cell-edge users may generate stronger levels of interference power to others, so they should be prioritized to emit lower levels of transmit power. To achieve soft frequency reuse in the uplink [39], one possible definition of $w_t^{(m,n)}$, for each pair of SBS $m \in \mathcal{M}$ and user $n \in \mathcal{N}_t^{(m)}$, can be

$$w_t^{(m,n)} = \sum_{m' \in \mathcal{M}, m' \neq m} \frac{H_t^{(n,m')}}{H_t^{(n,m)}}, \tag{2}$$

where each $H_t^{(n,m)}$ denotes propagation gain in the uplink from user $n$ to SBS $m$. If assuming the basic ground-to-air path loss model [34], to be discussed later, this definition of user weights can be distance-dependent and can be helpful to optimize transmission/interference distances while placing small cells. In the weighted sum of transmit power levels as in (1) and (2), the values of $\mathcal{N}_t^{(m)}$ and $w_t^{(m,n)}$ for $m \in \mathcal{M}$, $n \in \mathcal{N}_t^{(m)}$ are related to SCP and UCA, while the values of $\mathcal{K}_t^{(m,n)}$ and $p_t^{(m,n,k)}$ for $m \in \mathcal{M}$, $n \in \mathcal{N}_t^{(m)}$, $k \in \mathcal{K}_t^{(m,n)}$ are related to CRA and TPC.

There are a number of constraints for the minimization of objective function. First, each user should be associated with no more than one small cell in a certain stage. Note that each small cell can still accommodate multiple users at a time. Thus, UCA needs to satisfy

$$\sum_{m \in \mathcal{M}} a_t^{(m,n)} \leq 1, \quad \text{for } n \in \mathcal{N}. \tag{3}$$

Second, within a certain small cell, each channel should be allocated to no more than one user in a certain stage, to avoid intra-cell interference in the uplink. Note that each user can still take multiple channels at a time, and each channel can still be shared among multiple users in different cells. Thus, CRA needs to satisfy

$$\sum_{n \in \mathcal{N}_t^{(m)}} b_t^{(m,n,k)} \leq 1, \quad \text{for } m \in \mathcal{M},\ k \in \mathcal{K}. \tag{4}$$

Third, for a certain user in a certain small cell, transmit power over each channel being taken should ensure required signal-to-interference-plus-noise ratio (SINR). Thus, TPC needs to satisfy

$$\frac{p_t^{(m,n,k)} H_t^{(n,m)}}{I_t^{(m,k)} + Z_0^{(k)}} \geq \gamma^{(n)}, \quad \text{for } m \in \mathcal{M}, n \in \mathcal{N}_t^{(m)}, k \in \mathcal{K}_t^{(m,n)}, \tag{5}$$

where each $\gamma^{(n)}$ denotes user $n$'s requirement of SINR for successful uplink transmissions; each $I_t^{(m,k)}$ denotes aggregate interference power on channel $k$ locally observed at SBS $m$ in stage $t$, which is generated from co-channel users in neighboring cells; each $Z_0^{(k)}$ denotes average noise power on channel $k$. Fourth, when CRA and TPC are jointly considered, each user within a certain small cell should take sufficient number of channels to achieve required aggregate capacity (per unit bandwidth). Thus, CRA and TPC need to satisfy

$$\sum_{k \in \mathcal{K}_t^{(m,n)}} \log\left(1 + \frac{p_t^{(m,n,k)} H_t^{(n,m)}}{I_t^{(m,k)} + Z_0^{(k)}}\right) \geq \delta^{(n)}, \quad \text{for } m \in \mathcal{M}, n \in \mathcal{N}_t^{(m)}, \tag{6}$$

where each $\delta^{(n)}$ denotes user $n$'s requirement of capacity for satisfactory uplink transmissions, which can be collected as user $n$ requests channel access for an application.

In summary, the problem per stage for $t = 1, 2, \ldots, T$ that corresponds to each stage of the joint optimization of small cell deployment and uplink resource allocation is defined as

$$\begin{aligned} \mathbf{P}_0: \ & \text{find}: \ L_t, A_t, B_t, P_t; \\ & \text{minimize}: \ U_t = \sum_{m \in \mathcal{M}} U_t^{(m)}; \\ & \text{s.t.}: \ (3), (4), (5), (6). \end{aligned}$$

The formulated problem is a mixed-integer non-linear program (MINLP), which is NP-hard in general [39]. There are four sets of decision variables to determine, so the solution space can be prohibitively large. The problem per stage is complex to solve even assuming perfect global knowledge. Furthermore, it is especially challenging to keep refining the solution to

this problem stage by stage under user uncertainty. Instead of solving it in a centralized way, we can decouple its two subproblems—small cell deployment and uplink resource allocation, and let both tiers of macro- and small cells alternately contribute to problem solving.

## IV. Hybrid Multi-Agent Approach

In this section, we introduce the basic idea of our solution approach, and its two tiers of decision making will be elaborated in the next two sections.

To make the original problem tractable, we propose a hybrid multi-agent approach that follows a principal-agent model [40], in which the MBS (as principal) learns user demands from past stages and guides the SBSs (as agents) to take their part of responsibility for problem solving. Specifically, our approach takes advantage of both the MBS's capability of online learning and each SBS's potential of autonomous governance. Once small cell deployment and uplink resource allocation are decoupled, the first-tier MBS and the second-tier SBSs can undertake these two subproblems, respectively. On the one hand, the MBS should be in charge of the first subproblem. It is important to solve SCP and UCA based on a global picture of all small cells. On the other hand, given a solution to small cell deployment, each SBS can be responsible for its own part of the second subproblem. Within each small cell, it is possible to solve CRA and TPC only based on local knowledge of other small cells [39].

Our proposed approach operates with two tiers of sequential decision making under user uncertainty, requiring the leading MBS and the following SBSs to act alternately and evolve in response to each other. In absence of a priori user knowledge, reinforcement-learning-based models can be useful. On the first tier, the MBS interacts with its environment (including the SBSs), and refines its global decision according to the environment's feedback rewards. Similarly on the second tier, each SBS interacts with its environment (including the MBS), and refines its local decision according to the environment's state changes. The basic idea of our approach is illustrated in Fig. 1. More specifically, according to the decision feedbacks in past stages, the MBS refines its strategy for (multi-cell) SCP and UCA to optimize the global objective. Meanwhile, the MBS updates user knowledge through stage-by-stage reinforcement learning. We consider an adversarial bandit model to derive a mixed strategy that allows online exploration and exploitation. In response, each SBS adapts its strategy for (single-cell) CRA and TPC to optimize its local utility, a component of the global objective. Under the guidance of the MBS, all the SBSs participate in spectrum sharing through stage-by-stage multi-agent reinforcement learning. Because of the conflict of interest among coexisting small cells, we frame a stochastic game model to obtain an equilibrium solution. Generally in our hybrid approach, the first-tier MBS makes a sequence of coarse-grained global decisions and leads the group of all the second-tier SBSs, each of which makes a sequence of fine-grained local decisions. As a result, all the entities make their own contributions to solving the original problem. Our approach can be responsive to initially unknown and varying demands of user devices, and can be applicable to a resource-constrained cellular system, despite a modest loss of optimality.

## V. Small Cell Deployment

In this section, we focus on the refinement of the MBS's strategy for SCP and UCA. If considering the SBSs as part of the environment, the MBS can learn user demands and refine small cell deployment every time receiving a decision feedback from such an environment. The feedback for a MBS's decision is actually concluded by its subsequent SBSs' decisions thus is non-stochastic. Then, solving the subproblem of small cell deployment stage by stage can be viewed as playing with an adversarial or non-stochastic bandit [41]. In this model, one "arm" corresponds to one of the MBS's possible actions, but its reward distribution depends on the SBSs' responding actions and is thus not specifically known. After certain stages of trying (i.e., pulling arms), learning (i.e., updating reward beliefs), and adapting, the best global strategy of the MBS is attainable, which is capable of addressing the exploration-exploitation trade-off.

### A. Adversarial Bandit

For a finite time horizon, i.e., for $t = 1, 2, \ldots, T$, the MBS makes sequential decisions on small cell deployment by dealing with an adversarial bandit problem, which can be defined by a 2-tuple $< \mathcal{L}^M \times \mathcal{A}^{M\times N}, \bar{R}_\cdot(\cdot) >$: $\mathcal{L}^M \times \mathcal{A}^{M\times N}$ is a set of possible actions for the MBS, who takes an action $(L_t, A_t) \in \mathcal{L}^M \times \mathcal{A}^{M\times N}$ in stage $t$; $\bar{R}_t(L_t, A_t)$ is a feedback reward/cost for the MBS, which evaluates the applied action $(L_t, A_t)$ in stage $t$.

In each stage $t$, the MBS selects an action $(L_t, A_t) = (L^j, A^j)$ from a probability distribution $\pi_t$ over $\mathcal{L}^M \times \mathcal{A}^{M\times N} = \{(L^j, A^j) \,|\, j = 1, 2, \ldots, |\mathcal{L}^M \times \mathcal{A}^{M\times N}|\}$, in which each possible action $(L^j, A^j)$ is assigned with a probability $\pi_t(L^j, A^j)$. It is required to ensure that $\sum_j \pi_t(L^j, A^j) = 1$. The reward/cost function $\bar{R}_t$ needs to be defined to favor the strategy $\pi_t$ that minimizes the objective $U_t$ in the problem $\mathbf{P}_0$ on average. Hence, we can have

$$\bar{R}_t(L_t, A_t) = U_t(L_t, A_t, B_t^*, P_t^*), \tag{7}$$

which is an evaluation of not only the MBS's applied action $(L_t, A_t)$ but also the SBSs' best joint action $(B_t^*, P_t^*)$ in response (hidden by the environment for the MBS). In this way, minimizing the reward or actually the cost for spectrum sharing is equivalent to optimizing the global objective.

In the finite time horizon, a mixed strategy of the MBS, say $\pi = \{\pi_t \,|\, t = 1, 2, \ldots, T\}$, is able to generate a sequence of applied actions $(L_1, A_1), \ldots, (L_T, A_T)$. Comparing those $\pi$-chosen actions with the cost-minimizing one, the optimality of $\pi$ is evaluated with respect to the criterion of either expected regret $\bar{G}^\pi$ or pseudo-regret $\tilde{G}^\pi$, which can be defined by

$$\bar{G}^\pi = \mathbb{E}\left[\sum_{t=1}^{T} \bar{R}_t(L_t, A_t) - \min_j \sum_{t=1}^{T} \bar{R}_t(L^j, A^j)\right], \tag{8}$$

$$\tilde{G}^\pi = \mathbb{E}\left[\sum_{t=1}^{T} \bar{R}_t(L_t, A_t)\right] - \min_j \mathbb{E}\left[\sum_{t=1}^{T} \bar{R}_t(L^j, A^j)\right]. \tag{9}$$

The final goal of the MBS is to derive a mixed strategy $\pi^*$ for small cell deployment such that

$$\pi^* = \arg\min_{\pi} \bar{G}^{\pi} \text{ or } \arg\min_{\pi} \tilde{G}^{\pi}, \tag{10}$$

supposing the best responses of the SBSs behind the environment for uplink resource allocation.

### B. Subproblem Solution

It is still challenging to solve the adversarial bandit problem due to the huge size of the MBS's action space $\mathcal{L}^M \times \mathcal{A}^{M\times N}$. Fortunately, the two issues of SCP and UCA for small cell deployment can further be decoupled. Particularly, the MBS takes a primary action $A_t \in \mathcal{A}^{M\times N}$ for UCA in every stage $t$, i.e., to group $N$ users into $M$ clusters and associate each user cluster with a small cell. Based on $A_t$, the MBS takes a secondary action $L_t \in \mathcal{L}^M$ for SCP, i.e., to place each small cell above its associated user cluster. It is beneficial to address UCA before SCP (and subsequent CRA and TPC), because the involvement of binary variables for UCA gives a finite number of possible actions (arms) for the MBS's sequential decision making. Furthermore, each possible action $A_t$ leads the other subsequent decisions within a stage, and can be evaluated by a feedback cost $\bar{R}_t(A_t)$. Then, the action space can be downsized to $\mathcal{A}^{M\times N}$.

Before getting to find the optimal strategy or action of $A_t$ for UCA in stage $t$, we put initial focus on deriving the optimal action of $L_t$ for SCP assuming that $A_t$ is given. Note that while handling the issue of SCP, the MBS does not directly manage the subsequent issues of CRA and TPC, which have become local to the SBSs after problem decoupling. In the phase of SCP without knowing next decisions on CRA and TPC, the MBS has to coarsely optimize the coverage of every small cell, i.e., to optimize a part of the objective function in (1), and help the SBSs further optimize the other part of it. The idea here is that the transmission and interference distances in the uplink can be minimized and maximized, respectively, through the optimal placement of small cells. This can be an indirect way of improving soft frequency reuse in the uplink. Thus for SCP, the objective function in (1) is truncated to

$$V_t^{(m)} = \sum_{n\in\mathcal{N}_t^{(m)}} \lambda_t^{(n)} w_t^{(m,n)}, \tag{11}$$

in which the values of $\lambda_t^{(n)}$ for $n \in \mathcal{N}$ can be estimated stage by stage; the values of $\mathcal{N}_t^{(m)}$ and $w_t^{(m,n)}$ for $m \in \mathcal{M}$, $n \in \mathcal{N}_t^{(m)}$ are determined by $A_t$ and $L_t$, respectively. The optimization of transmission/interference distances can be achieved by properly defining $w_t^{(m,n)}$ for each pair of SBS $m$ and user $n$. If assuming the basic ground-to-air path loss model [34], each value of propagation gain $H_t^{(n,m)}$ in (2) is still inversely proportional to the Euclidean distance $d_t^{(m,n)} = ||l_t^{(m)} - l^{(n)}||$ (to the power of path loss exponent $\alpha$), between SBS $m$ at $l_t^{(m)} \in \mathcal{L}$ and user $n$ at $l^{(n)} \in \mathcal{L}$. Then, each user weight can become a ratio of transmission/interference distances. Alternatively, each user weight can be directly defined like this, instead of following (2). Hence, we can rewrite the above $V_t^{(m)}$ as

$$V_t^{(m)} = \sum_{n\in\mathcal{N}_t^{(m)}} \lambda_t^{(n)} (d_t^{(m,n)})^{\alpha} \sum_{m'\in\mathcal{M}, m'\neq m} (d_t^{(m',n)})^{-\alpha}. \tag{12}$$

Then, according to the original problem $\mathbf{P}_0$ of joint optimization, the subsubproblem of SCP per stage can be defined as

$$\begin{aligned} \mathbf{P}_{SCP}: \ & \text{given}: \ A_t; \\ & \text{find}: \ L_t; \\ & \text{minimize}: \ V_t = \sum_{m\in\mathcal{M}} V_t^{(m)}; \\ & \text{s.t.}: \ l_t^{(m)} \in \mathcal{L}^{(m)} form \in \mathcal{M}. \end{aligned}$$

Each SBS $m$ only searches a location space $\mathcal{L}^{(m)} \subset \mathcal{L}$ that is local to its associated users $\mathcal{N}_t^{(m)}$. According to optimization theory, we can show that such a subsubproblem is directly solvable.

*Lemma 1:* Given that in each stage $t = 1, 2, \ldots, T$, the values of $\mathcal{N}_t^{(m)}$ for $m \in \mathcal{M}$ are fixed by a certain $A_t$, and the values of $w_t^{(m,n)}$ for $m \in \mathcal{M}$, $n \in \mathcal{N}_t^{(m)}$ are defined by (2), and each $\mathcal{L}^{(m)}$ for $m \in \mathcal{M}$ is closed and convex, then there exists the optimal solution $L_t^*$ to $\mathbf{P}_{SCP}$.

*Proof:* The distance $d_t^{(\bar{m},\bar{n})}$ between any SBS $\bar{m} \in \mathcal{M}$ and any user $\bar{n} \in \mathcal{N}$ only appears once in $V_t$. Specifically, if user $\bar{n}$ is associated with SBS $\bar{m}$, their distance is considered as a transmission distance in the part of $(d_t^{(m,n)})^{\alpha}$; if not, it is considered as an interference distance in the part of $(d_t^{(m',n)})^{-\alpha}$. We can see that the minimization of $V_t$ for SCP contributes to the maximization of inter-cell interference distances as well as the minimization of intra-cell transmission distances. If considering $V_t$ as a function of any $l_t^{(\bar{m})} = (x_t^{(\bar{m})}, y_t^{(\bar{m})}, z_t^{(\bar{m})}) \in \mathcal{L}^{(\bar{m})}$, the objective function becomes the following form:

$$\begin{aligned} V_t = & \sum_{n\in\mathcal{N}_t^{(\bar{m})}} \kappa_t^{(n)} (d_t^{(\bar{m},n)})^{\alpha} \\ & + \sum_{n\notin\mathcal{N}_t^{(\bar{m})}} \kappa_t^{(n)} (d_t^{(\bar{m},n)})^{-\alpha} + \kappa_t, \end{aligned} \tag{13}$$

where each $\kappa_t^{(n)}$ or $\kappa_t$ is a positive constant. The first part is to minimize transmission distances in cell $\bar{m}$, and the second part is to maximize interference distances from the other cells. In a small enough search space $\mathcal{L}^{(\bar{m})}$, the minimization of $V_t$ is dominated by the minimization of the first part, which is a convex function of $l_t^{(\bar{m})}$. A typical contour plot of $V_t$ with respect to $(x_t^{(\bar{m})}, y_t^{(\bar{m})})$ ($z_t^{(\bar{m})} = 10$ [32]) is illustrated in Fig. 2, when associated users $n \in \mathcal{N}_t^{(\bar{m})}$ are generated near this region's center and interfering users $n \notin \mathcal{N}_t^{(\bar{m})}$ are distributed outside the local search space. Hence, any SBS $\bar{m}$ can find its own optimal $l_t^{*(\bar{m})}$ that helps minimize $V_t$ while temporarily fixing the other SBSs. Through alternating minimization, the optimal solution $L_t^*$ that minimizes $V_t$ is achievable for a given $A_t$. □

Existing non-linear optimization solvers, such as "fmincon" in MATLAB, can be applied to optimize SCP subject to UCA. For the reason that $L_t^*$ (and subsequent $(B_t^*, P_t^*)$ for CRA and TPC) always follows $A_t$, the MBS's sequential decision

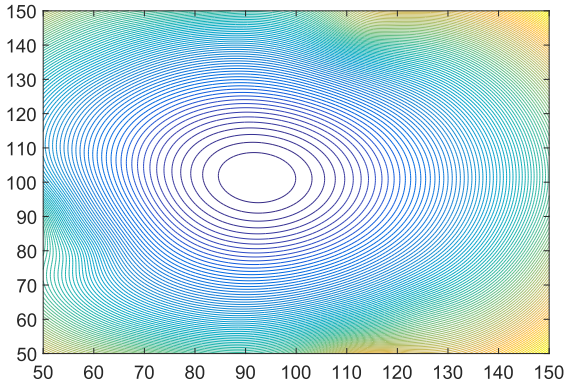


Fig. 2. An example of $V_t$ as a function of $(x_t^{(\bar{m})}, y_t^{(\bar{m})})$ (in meters).

making can only involve finding $A_t$ stage by stage. To make such decisions, the subsubproblem of UCA per stage can be defined as

$$\begin{aligned} \mathbf{P}_{UCA}: \ & \text{find}: \ A_t; \\ & \text{minimize}: \ U_t = \sum_{m\in\mathcal{M}} U_t^{(m)}; \\ & \text{s.t.}: \ (3). \end{aligned}$$

This subsubproblem is the one to be addressed as an adversarial bandit problem, in which the MBS's action space is revised to $\mathcal{A}^{M\times N}$. Because multiple users close to each other will probably be associated with the same small cell nearby, these users can be tied together for UCA and the action space can further be replaced with a (small) subset of it. Alternatively, the action space can be constructed based on the possibilities of grouping $N$ users into $M$ clusters.

Consider a finite action space for UCA, say $\mathcal{A}^J = \{A^j \mid j = 1, 2, \ldots, J\} \subseteq \mathcal{A}^{M\times N}$. In each stage $t$, the MBS selects an action $A_t = A^j$ from a probability distribution $\pi_t$ over $\mathcal{A}^J$, in which each possible action $A^j$ is assigned with a probability $\pi_t(A^j)$. It is also required to ensure that $\sum_j \pi_t(A^j) = 1$. In bandit theory, the strategy $\pi_t$ can be updated by applying the exponential-weight algorithm [42]. Specifically, if the MBS observes a cost $\bar{R}_t(A_t)$ after taking $A_t$, let

$$\omega_{t+1}(A^j) = \begin{cases} \omega_t(A^j)\exp(-\eta\dfrac{\bar{R}_t(A^j)}{\pi_t(A^j)}) & \text{if } A_t = A^j; \\ \omega_t(A^j) & \text{otherwise,} \end{cases} \quad (14)$$

where $\eta$ is a control factor, which will be discussed later; each $\omega_{t+1}(A^j)$ denotes action $A^j$'s weight for probabilistic selection in stage $t+1$, and is then normalized to

$$\pi_{t+1}(A^j) = \frac{\omega_{t+1}(A^j)}{\sum_{j=1}^{J}\omega_{t+1}(A^j)}. \quad (15)$$

In this way, the strategy $\pi_{t+1}$ to be utilized in stage $t+1$ can be derived based on only the currently observed cost $\bar{R}_t(A_t)$ by the end of stage $t$. Generally speaking, if the applied action $A_t = A^j$ returns a higher cost $\bar{R}_t(A^j)$, then this action $A^j$ will be less preferred in the future and will be assigned with a lower probability $\pi_{t+1}(A^j)$. The main steps are summarized in Algorithm 1. In the finite time horizon, we are able to show

**Algorithm 1** Small Cell Deployment at the MBS

1: **for** $t = 1, 2, \ldots, T$ **do**
2: **if** $t == 1$ **then**
3: **for** $j = 1, 2, \ldots, J$ **do**
4: **initialize** $\omega_t(A^j) = 1$ and $\pi_t(A^j) = \frac{1}{J}$
5: **end for**
6: **end if**
7: **select** $A_t^* = A^j$ from $\pi_t$ probabilistically
8: **obtain** $L_t^*$ according to $A_t^*$ via an optimization solver
9: **send** $(L_t^*, A_t^*)$ to the SBSs, and wait for their responding actions for uplink resource allocation
10: **receive** $(B_t^*, P_t^*)$ from the SBSs, which can be obtained according to $(L_t^*, A_t^*)$ via locally running Algorithm 2
11: **compute** $\bar{R}_t(A_t^* = A^j)$ according to $(L_t^*, A_t^*, B_t^*, P_t^*)$
12: **if** $t < T$ **then**
13: **for** $j = 1, 2, \ldots, J$ **do**
14: **update** $\omega_{t+1}(A^j)$ and $\pi_{t+1}(A^j)$ via (14) and (15)
15: **end for**
16: **end if**
17: **end for**

that this algorithm provides regret guarantees for UCA, but it is still heuristic due to the involvement of user uncertainty.

*Theorem 1:* Given that in a period of $T$ stages, the MBS runs Algorithm 1 to refine the probability distributions $\pi_t$ over $\mathcal{A}^J$ for $t = 1, 2, \ldots, T$, and generates a sequence of applied actions $A_1^*, \ldots, A_T^*$ accordingly, then there exists an upper bound on expected regret $\bar{G}^\pi$ or pseudo-regret $\tilde{G}^\pi$ in the finite-horizon adversarial bandit defined by $\mathbf{P}_{UCA}$.

*Proof:* It has been proven in [43] that the pseudo-regret achieved by the exponential-weight algorithm is upper bounded, and $\tilde{G}^\pi \le \sqrt{2\,TJ\ln J}$ holds for a certain setting of control factor $\eta$ in (14). Moreover, the environment responds to the MBS's current action $A_t^*$ with a cost $\bar{R}_t(A_t^*)$, which does not depend on the MBS's past actions $A_1^*, \ldots, A_{t-1}^*$, so the environment is oblivious. In this case, the pseudo-regret equals to the expected regret, which is thus also upper bounded, and $\bar{G}^\pi \le \sqrt{2\,TJ\ln J}$ also holds. □

In summary, the first subproblem can be solved by running Algorithm 1 at the MBS, which interacts with Algorithm 2 locally run at each SBS, to be discussed in the next section. The output of Algorithm 1 includes the MBS's strategy for (optimal) SCP and (heuristic) UCA along with a sequence of corresponding actions.

## VI. Uplink Resource Allocation

In this section, we study the refinement of each SBS's local strategy for CRA and TPC under the guidance of the MBS. If considering the MBS as part of the environment, each SBS makes a contribution to uplink resource allocation every time receiving a solution to small cell deployment from such an environment. Because of potential mutual interference among coexisting small cells, solving the subproblem of uplink resource allocation stage by stage can be viewed as playing with a stochastic or Markov game [40], [44]. In this model, each SBS acts as a "player" on behalf of its entire small

cell and may compete with others for spectrum sharing. The MBS's current action can be viewed as a system state that guides the SBSs' responding actions, which in return trigger the transition to a new state (determined by the MBS's next action). After certain stages of refinement, the local strategies of all the SBSs are expected to constitute an equilibrium in the stochastic game for uplink resource allocation.

### A. Stochastic Game

For a finite time horizon, i.e., for $t = 1, 2, \ldots, T$, the group of all the SBSs makes sequential joint decisions on uplink resource allocation by being involved in a stochastic game problem, which can be defined by a 5-tuple $< \mathcal{M}, \mathcal{L}^M \times \mathcal{A}^{M\times N}, \mathcal{B}^{N\times K} \times \mathcal{P}^{N\times K}, Q_\cdot(\cdot,\cdot,\cdot), R_\cdot^\cdot(\cdot,\cdot) >$: $\mathcal{M}$ is a set of game players, and each SBS $m \in \mathcal{M}$ plays for its own small cell; $\mathcal{L}^M \times \mathcal{A}^{M\times N}$ is a set of system states, and the leading action $(L_t^*, A_t^*) \in \mathcal{L}^M \times \mathcal{A}^{M\times N}$ gives the state in stage $t$; $\mathcal{B}^{N\times K} \times \mathcal{P}^{N\times K}$ is a set of possible actions for each SBS $m$, who takes an action $(B_t^{(m)}, P_t^{(m)}) \in \mathcal{B}^{N\times K} \times \mathcal{P}^{N\times K}$ in stage $t$, and equivalently the group of SBSs takes a joint action $(B_t, P_t) \in \mathcal{B}^{M\times N\times K} \times \mathcal{P}^{M\times N\times K}$ in stage $t$; $Q_t((L_t^*, A_t^*), (B_t, P_t), (L_{t+1}^*, A_{t+1}^*))$ is a state transition probability, which characterizes the likelihood that the state moves from $(L_t^*, A_t^*)$ in stage $t$ to $(L_{t+1}^*, A_{t+1}^*)$ in stage $t+1$ under the joint action $(B_t, P_t)$ taken in stage $t$; $R_t^{(m)}((L_t^*, A_t^*), (B_t, P_t))$ is a feedback payoff/cost for each SBS $m$, which evaluates the applied action $(B_t^{(m)}, P_t^{(m)})$ taken at the state $(L_t^*, A_t^*)$ in stage $t$.

In each stage $t$, each SBS $m$ observes the state $(L_t^*, A_t^*)$ and takes an action $(B_t^{(m)}, P_t^{(m)})$. The resulting joint action $(B_t, P_t)$ triggers the move to stage $t+1$ and the transition to state $(L_{t+1}^*, A_{t+1}^*)$. The function $Q_t$ can be estimated based on the MBS's mixed strategy. Let each

$$\begin{aligned} &Q_t((L_t^*, A_t^*), (B_t, P_t), (L_{t+1}^*, A_{t+1}^*)) \\ &\quad = \Pr\{(L_{t+1}^*, A_{t+1}^*) | (L_t^*, A_t^*), (B_t, P_t)\} \\ &\quad = \Pr\{A_{t+1}^* | (L_t^*, A_t^*), (B_t, P_t)\} \cdot \Pr\{L_{t+1}^* | A_{t+1}^*\}, \end{aligned} \quad (16)$$

where the probability $\Pr\{A_{t+1}^* | (L_t^*, A_t^*), (B_t, P_t)\}$ is given by $\pi_{t+1}$ for UCA, and the probability $\Pr\{L_{t+1}^* | A_{t+1}^*\}$ depends on how the non-linear optimization for SCP is solved and can be estimated empirically if $L_{t+1}^*$ is not unique to $A_{t+1}^*$. The payoff/cost function $R_t^{(m)}$ for each SBS $m$ should be consistent with the reward/cost function $\bar{R}_t$ for the MBS and be a component of the global objective. Hence, let each

$$R_t^{(m)}((L_t^*, A_t^*), (B_t, P_t)) = U_t^{(m)}(L_t^*, A_t^*, B_t^{(m)}, P_t^{(m)}), \quad (17)$$

which is a measure of SBS $m$'s applied action $(B_t^{(m)}, P_t^{(m)})$ in response to the MBS's leading action $(L_t^*, A_t^*)$.

In the finite time horizon, a pure strategy of each SBS $m$, say $\theta^{(m)} = \{\theta_t^{(m)} \,|\, t = 1, 2, \ldots, T\}$, needs to be refined, in which let each $\theta_t^{(m)}(L_t^*, A_t^*) = (B_t^{(m)}, P_t^{(m)})$. Given the joint strategy of all the SBSs except SBS $m$, say $\boldsymbol{\theta}^{(-m)}$, the optimality of $\theta^{(m)}$ can be evaluated with respect to expected total discounted payoff/cost, which is defined by

$$G^{\{\theta^{(m)}, \boldsymbol{\theta}^{(-m)}\}} = \mathbb{E}\left[\sum_{t=1}^{T} \beta^{t-1} R_t^{(m)}(, (B_t, P_t))\right], \quad (18)$$

where $\beta \in (0, 1)$ is a discount factor used to put more focus on more recent stages. The goal of SBS $m$ is to derive a greedy strategy $\theta^{*(m)}$ for (local) uplink resource allocation such that

$$\theta^{*(m)} = \arg\min_{\theta^{(m)}} G^{\{\theta^{(m)}, \boldsymbol{\theta}^{(-m)}\}}. \quad (19)$$

More importantly, the set of all the SBSs' greedy strategies, $\boldsymbol{\theta}^* = \{\theta^{*(m)} \,|\, m \in \mathcal{M}\}$, should be able to establish an equilibrium in the stochastic game.

### B. Subproblem Solution

It is still challenging to guarantee that the SBSs can agree on a common set of greedy strategies as an equilibrium solution to the stochastic game problem. Due to the involvement of binary variables for CRA like those for UCA, the two issues of CRA and TPC for uplink resource allocation can further be decoupled as well. Particularly, each SBS $m$ takes a primary action $B_t^{(m)} \in \mathcal{B}^{N\times K}$ for CRA in every stage $t$. Based on the resulting $B_t$, each SBS $m$ takes a secondary action $P_t^{(m)} \in \mathcal{P}^{N\times K}$ for TPC. Hence, each stage game that is a non-cooperative game to be played repeatedly can be decoupled into two subgames, for CRA and TPC respectively.

We first study one stage game in $t$. Before getting to find the equilibrium solution of $B_t$ to the subgame for CRA in stage $t$, we initially focus on deriving the optimal joint solution of $P_t$ to the subgame for TPC assuming that $B_t$ is given. According to the original problem $\mathbf{P}_0$, the subsubproblem of TPC per stage to be locally solved by each SBS $m$ can be defined as

$$\begin{aligned} \mathbf{P}_{TPC}: \ &\text{given}: \ L_t^*, A_t^*, B_t, P_t^{(-m)}; \\ &\text{find}: \ P_t^{(m)}; \\ &\text{minimize}: \ U_t^{(m)}; \\ &\text{s.t.}: \ (5), (6). \end{aligned}$$

Each SBS $m$ may solve this subsubproblem multiple times before its $P_t^{(m)}$ and others' $P_t^{(-m)}$ establish an equilibrium. According to the fixed point theorem in game theory [39], we can show that the subgame problem defined by such a subsubproblem is uniquely solvable.

*Lemma 2:* Given that in each stage $t = 1, 2, \ldots, T$, the values of $\mathcal{N}_t^{(m)}$ for $m \in \mathcal{M}$ and $w_t^{(m,n)}$ for $m \in \mathcal{M}$, $n \in \mathcal{N}_t^{(m)}$ are determined by $(L_t^*, A_t^*)$, and the values of $\mathcal{K}_t^{(m,n)}$ for $m \in \mathcal{M}$, $n \in \mathcal{N}_t^{(m)}$ are fixed by a certain $B_t$, and $\mathcal{P}$ is closed and convex, then there exists the (unique) Nash equilibrium $P_t^*$ in the subgame defined by $\mathbf{P}_{TPC}$.

*Proof:* The global objective is the weighted sum of transmit power levels. If every user lowers down its transmit power (also interference power to others) on each channel, all the co-channel users would experience the minimum levels of interference power and finally achieve their minimum levels of transmit power when the equalities in (5) hold for basic SINR guarantees. According to [39], the resulting set of all users' minimum power levels gives the optimal joint solution $P_t^*$, in which each $P_t^{*(m)}$ solves $\mathbf{P}_{TPC}$ for one SBS $m$. Furthermore, the solution $P_t^*$ to the system of equations from (5) is shown to be unique. Following the same logic as the proof for Theorem 1 in [39], we can prove that the

fixed point theorem holds for the subgame for TPC. Hence, the Nash equilibrium can be established by the unique $P_t^*$ for a given $B_t$. □

Some existing distributed algorithms can optimize TPC subject to CRA, such as iterative water-filling algorithm. Because $P_t^*$ always follows $B_t$, we now focus on the subgame for CRA. Each SBS $m$ locally solves the subsubproblem of CRA per stage, which can be defined as

$$\begin{aligned}\mathbf{P}_{CRA}: \ &\text{given}: \ L_t^*, A_t^*, B_t^{(-m)};\\ &\text{find}: \ B_t^{(m)};\\ &\text{minimize}: \ U_t^{(m)};\\ &\text{s.t.}: \ (4),(5),(6).\end{aligned}$$

Each SBS $m$ may solve this subsubproblem multiple times before its $B_t^{(m)}$ and others' $B_t^{(-m)}$ are commonly agreed upon. If the subgame problem defined by such a subsubproblem can be solved in a distributed manner, we can show that the stage game in $t$ converges to a Nash equilibrium.

In each stage $t$, each SBS $m$ should locally make a binary decision $b_t^{(m,n,k)}$ for each pair of user $n \in \mathcal{N}_t^{(m)}$ and channel $k \in \mathcal{K}$. The greedy SBSs may not reach a commonly agreed solution of $B_t$ due to potential mutual interference. Even without timely coordination among the SBSs, however, we can still guarantee the convergence of the subgame for CRA through a distributed algorithm for uplink spectrum reuse [39]. Specifically in each small cell $m$, channel allocation for each user $n$ depends on its weight $w_t^{(m,n)}$. According to the definition of user weights in (2), the users with larger weights are either more likely to cause interference to others or more vulnerable to interference from others. Hence, such users should be prioritized to get high-quality channels. For channel quality evaluation, each SBS $m$ can locally observe the aggregate interference power $I_t^{(m,k)}$ on each channel $k$, so the good channels can be dedicated ones or shared ones with low interference levels. Then based on both user weights and interference measurements, each SBS can take its associated users in turns, each of which should be allocated with certain channels subject to the constraints (4), (5), (6). Every user needs to emit sufficiently high (but not so high) level of transmit power on each taken channel for acceptable SINR, and also needs to get sufficient number of channels for desired aggregate capacity. Given that the equalities in (5) hold after the adaptation of power levels, then to further ensure (6), the number of channels to be allocated to each user $n \in \mathcal{N}$ can be computed by

$$K^{(n)} = \left\lceil \frac{\delta^{(n)}}{\log(1+\gamma^{(n)})} \right\rceil. \tag{20}$$

As summarized in Algorithm 2, each SBS $m$ obtains a solution of $B_t^{(m)}$ heuristically, and may further refine it in response to the joint solution of $B_t^{(-m)}$ from other SBSs.

To guarantee the convergence of the subgame for CRA, we rewrite each local objective $U_t^{(m)}$ by adding a cost function

$$C_t^{(m)} = \varphi^{(m)} I_t^{(m)} = \varphi^{(m)} \sum_{k\in\mathcal{K}} I_t^{(m,k)}, \tag{21}$$

**Algorithm 2** Uplink Resource Allocation at Each SBS $m$

**for** $t = 1, 2, \ldots, T$ **do**
  **wait** for the leading action for small cell deployment
  **receive** $(L_t^*, A_t^*)$ from the MBS, which can be obtained via running Algorithm 1
  **initialize** $B_t^{(m)}$ to a certain solution and $\varphi^{(m)}$ to a positive constant $\varphi_0^{(m)}$, and play the subgame for CRA
  **repeat**
    **wait** for the next turn to refine $B_t^{(m)}$, while others are taking turns to refine $B_t^{(-m)}$
    **obtain** $P_t^{(m)}$ according to the current $B_t^{(m)}$ in the subgame for TPC, and measure $I_t^{(m,k)}$ for $k \in \mathcal{K}$
    **compute** $R_t^{(m)}$ acc. to the current $(L_t^*, A_t^*, B_t^{(m)}, P_t^{(m)})$
    **set** $U_0^{(m)} = R_t^{(m)}$, $I_0^{(m)} = \sum_{k\in\mathcal{K}} I_t^{(m,k)}$, $\hat{U}_0^{(m)} = U_0^{(m)} + \varphi_0^{(m)} I_0^{(m)}$ accordingly, and set $B_0^{(m)} = B_t^{(m)}$
    **sort** the sequence of users $n \in \mathcal{N}_t^{(m)}$ by their values of $w_t^{(m,n)}$ in descending order, and let the sorted sequence of users be $\vec{\mathcal{N}} = \{n_1, n_2, \ldots, n_{N_t^{(m)}}\}$
    **reset** $B_t^{(m)} = \mathbf{0}$, and measure $I_t^{(m,k)}$ for $k \in \mathcal{K}$
    **sort** the sequence of channels $k \in \mathcal{K}$ by their values of $I_t^{(m,k)}$ in ascending order, and let the sorted sequence of channels be $\vec{\mathcal{K}} = \{k_1, k_2, \ldots, k_K\}$
    **for** $i = 1, 2, \ldots, N_t^{(m)}$ **do**
      **for** $j = 1, 2, \ldots, K$ **do**
        **if** $\sum_{i'=1}^{N_t^{(m)}} b_t^{(m,n_{i'},k_j)} == 0$ **then**
          **set** $b_t^{(m,n_i,k_j)} = 1, \ldots, b_t^{(m,n_i,k_j+K^{(n)}-1)} = 1$
        **end if**
      **end for**
    **end for**
    **obtain** $P_t^{(m)}$ according to the updated $B_t^{(m)}$ in the subgame for TPC, and measure $I_t^{(m,k)}$ for $k \in \mathcal{K}$
    **compute** $R_t^{(m)}$ acc. to the updated $(L_t^*, A_t^*, B_t^{(m)}, P_t^{(m)})$
    **set** $U^{(m)} = R_t^{(m)}$, $I^{(m)} = \sum_{k\in\mathcal{K}} I_t^{(m,k)}$, $\hat{U}^{(m)} = U^{(m)} + \varphi^{(m)} I^{(m)}$ accordingly
    **if** $\hat{U}^{(m)} \geq \hat{U}_0^{(m)}$ **then**
      **set** $B_t^{(m)} = B_0^{(m)}$ (give up this updated $B_t^{(m)}$)
    **end if**
    **if** $!(U^{(m)} \leq U_0^{(m)}$ && $I^{(m)} \leq I_0^{(m)})$ **then**
      **increase** $\varphi^{(m)}$ by a positive constant $\varphi_c^{(m)}$
    **end if**
  **until** the end of stage $t$
**end for**

where $\varphi^{(m)}$ is a positive price factor used to "punish" SBS $m$ if its strategy $B_t^{(m)}$ does not contribute to the desired equilibrium. More details are described in Algorithm 2. According to the Lyapunov's direct stability theorem in non-linear control theory [40], [45], we can show that the convergence of the stage game in $t$ for CRA and TPC to an equilibrium can be guaranteed.

*Lemma 3:* Given that in each stage $t = 1, 2, \ldots, T$, all the SBSs run Algorithm 2 independently to act in response to $(L_t^*, A_t^*)$, then there exists a Nash equilibrium $(B_t^*, P_t^*)$ in the stage game defined by $\mathbf{P}_{CRA}$ and $\mathbf{P}_{TPC}$.

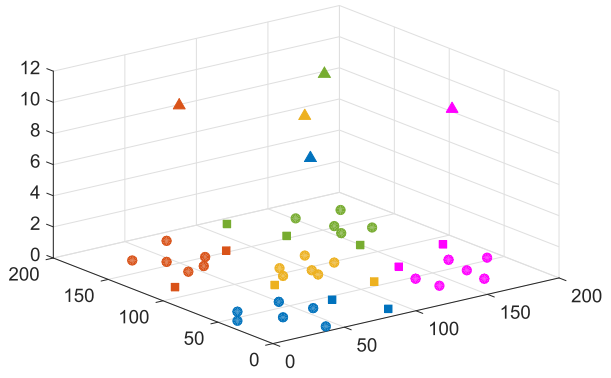


Fig. 3. An example of simulation scenario (in meters): triangles represent UAV-mounted SBSs; circles and squares represent user devices; and colors represent user-cell associations.

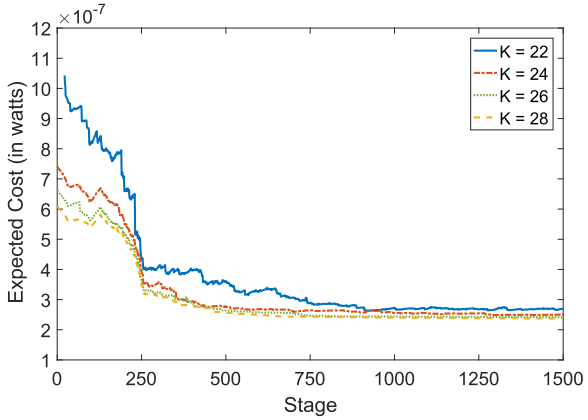


Fig. 4. Convergence behavior of our approach ($T = 2500$, $\eta_0 = 1 \times 10^8$).

*Proof:* The adaptation of price factor in Algorithm 2 makes $\varphi^{(m)}$ be non-decreasing and $\hat{U}_0^{(m)}$ be decreasing. Once $\varphi^{(m)}$ has been increased to always make $\hat{U}^{(m)} \geq \hat{U}_0^{(m)}$ hold, SBS $m$ no longer changes its strategy $B_t^{*(m)}$ in the subgame for CRA, which then gives the minimum $\hat{U}_0^{*(m)}$. Once all the SBSs give up acting in the subgame for CRA, a certain common $B_t^*$ is agreed upon. After that, the unique $P_t^*$ is obtained in the subgame for TPC under the fixed $B_t^*$. Following the same logic as the proof for Theorem 2 in [39], we can prove that the converged solution $(B_t^*, P_t^*)$ is a critical point and $\sum_{m\in\mathcal{M}}(\hat{U}_0^{(m)} - \hat{U}_0^{*(m)})$ is a Lyapunov function, so the Lyapunov's direct stability theorem holds for the stage game. Therefore, a Nash equilibrium can be established by the stable $(B_t^*, P_t^*)$ for a given $(L_t^*, A_t^*)$. □

After the analysis of one stage game in $t$, we study the multi-stage stochastic game that includes a sequence of such stage games for $t = 1, 2, \ldots, T$. In the finite time horizon, we are able to prove the existence of an equilibrium solution.

*Theorem 2:* Given that in a period of $T$ stages, the MBS runs Algorithm 1 to control the state transitions from $(L_t^*, A_t^*)$ to $(L_{t+1}^*, A_{t+1}^*)$ for $t = 1, 2, \ldots, T-1$, and all the SBSs run Algorithm 2 independently to generate a sequence of Nash equilibria $(B_1^*, P_1^*), \ldots, (B_T^*, P_T^*)$ accordingly, then there exists a Markov-perfect equilibrium $\boldsymbol{\theta}^*$ in the finite-horizon stochastic game defined by $\mathbf{P}_{CRA}$ and $\mathbf{P}_{TPC}$.

*Proof:* According to game theory, a finite-horizon multi-stage game always has a subgame-perfect equilibrium as long as each of its stage games has a Nash equilibrium. Therefore, a Markov-perfect equilibrium $\boldsymbol{\theta}^*$, i.e., a subgame-perfect equilibrium in Markov strategies, can be established by $(B_1^*, P_1^*), \ldots, (B_T^*, P_T^*)$ in this stochastic game [40]. □

In summary, the second subproblem can be solved by running Algorithm 2 locally at each SBS, which interacts with Algorithm 1 run at the MBS. The output of Algorithm 2 is the SBS's local strategy for (heuristic) CRA and (optimal) TPC as part of an equilibrium solution.

## VII. Performance Evaluation

In this section, we evaluate our hybrid multi-agent approach by simulations in MATLAB. Specifically, we consider an outdoor hotspot with the size of $200 \times 200$ in meters, in which $M$ UAV-mounted small cells are deployed to serve $N$ randomly distributed user devices by sharing $K$ available channels. An example of it is shown in Fig. 3, where $M = 5$ and $N = 40$. In each stage $t = 1, 2, \ldots, T$, each user $n \in \mathcal{N}$ transmits uplink data with a constant yet hidden probability $\lambda^{(n)}$, which can be estimated based on user $n$'s frequency of uplink transmissions by stage $t$. In Algorithm 1, the action space for UCA, $\mathcal{A}^J$, should not be too large due to the online feature of our approach. To get a tractable value of $J$, for example in Fig. 3, the potentially "inner" users for a cell (circles with the same color) can be associated with the same SBS, and each of the potentially "edge" users for any cell (squares with changeable colors) can be associated with one of the nearest SBSs. Using such a heuristic, we can make $J = 256$ for this example with a modest loss of optimality. For each SBS $m \in \mathcal{M}$, its local search space for SCP, $\mathcal{L}^{(m)} \subset \mathcal{L}$, is limited to a convex 3D space above its associated user cluster $\mathcal{N}_t^{(m)}$ on the ground, and is lower bounded by the minimum altitude of 10 meters [32]. We set the control factor $\eta = \eta_0\sqrt{\frac{2\ln J}{TJ}}$ in Algorithm 1 [43], where $\eta_0$ is a scale factor for $\bar{R}_t$ in (14). In Algorithm 2, the search space for TPC, $\mathcal{P}$, is a convex space that is upper bounded by the maximum transmit power of 23 dBm. We set other parameters in Algorithm 2 as follows: the threshold of SINR for each user $n$, $\gamma^{(n)} = 15$ dB; the number of channels for each user $n$, $K^{(n)} = 1$; the path loss exponent $\alpha = 3$; the average noise power on each channel $k$, $Z_0^{(k)} = -104$ dBm; the price factor $\varphi_0^{(m)} = \varphi_c^{(m)} = 1$. Now we evaluate our approach on four aspects.

### A. Convergence Behavior

Our approach involves two tiers of decision-making entities, so Algorithms 1 and 2 need to coordinate efficiently and converge eventually to a final joint strategy. To clearly show convergence behavior, we can keep track of the objective value achieved by our approach. In each stage $t$, a leading action $A_t = A^j$ is chosen from a probability distribution $\pi_t$ over $\mathcal{A}^J$, and a feedback cost $\bar{R}_t = U_t$ is received thereafter. Hence, we can compute the expected cost value, say $\bar{U}_t = \sum_j U_T^*(A^j)\pi_t(A^j)$, where each $U_T^*(A^j)$ records the

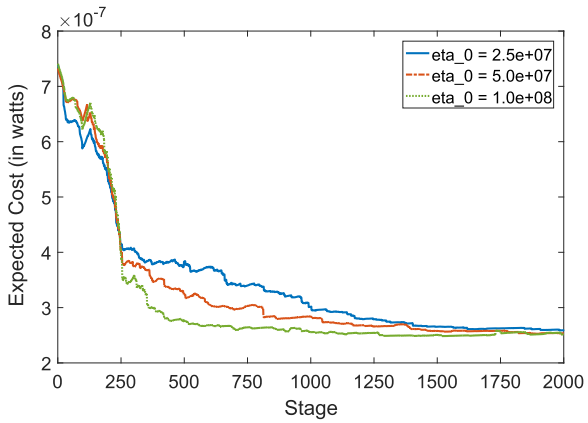


Fig. 5. Impact of $\eta$ on convergence behavior ($T = 2500$, $K = 24$).

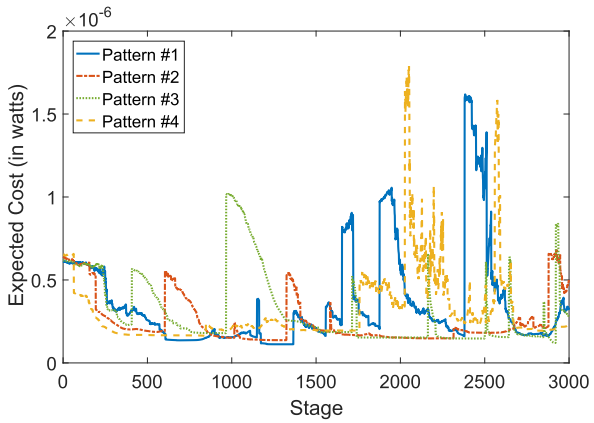


Fig. 6. Convergence behavior of our approach under user mobility ($T = 3000$, $K = 28$, $\eta_0 = 1 \times 10^8$).

minimum cost for a possible action $A^j \in \mathcal{A}^J$ by stage $T$. Given a sufficiently large value of $T$, the typical changing process of $\bar{U}_t$ is exemplified in Fig. 4. We set $T = 2500$, but our approach runs fast in each stage thanks to its low complexity. We can see that as user knowledge is accumulated after an initial evaluation of all possible actions, the refinement of $\pi_t$ gradually stabilizes $\bar{U}_t$ given a value of $K$. Despite minor probabilistic fluctuations, our approach can steadily converge to a certain level of objective value. The online capability of our approach allows it to be terminated anytime. Furthermore, we adjust the value of $\eta$ or $\eta_0$ to analyze its impact on convergence progress, which is shown in Fig. 5. We can see that a larger $\eta$ gives a faster convergence process, while a smaller $\eta$ gives a slower convergence process. In the following, we set $\eta_0 = 1 \times 10^8$. Overall, it is suggested that the convergence behavior of our approach is satisfactory.

Even though common wireless devices in an outdoor hotspot, e.g., event spectators' mobile devices or regular sensor nodes, do not change their locations significantly while being served, we still evaluate the impact of low-level user mobility on the convergence of our approach. A random waypoint model is used to generate user mobility patterns. Each user occasionally moves towards a random destination with a random walking-level speed from 0 to 3 meters/stage. The pause time of a user is randomly chosen from 300 to 500 stages. Like above, the convergence process in terms of $\bar{U}_t$ is exemplified in Fig. 6. We can see that while the action space for UCA keeps temporarily unchanged, our approach converges as usual despite small user movements in nearly every stage. Once the action space has to change, our approach needs to redo its strategy refinement. Here the best actions in the past are recorded, so that they can be assigned with high selection probabilities (if still in the action space) to accelerate the convergence process. Therefore, our approach can support a certain level of user mobility.

### *B. Strategy Correctness*

Our approach is driven by the MBS's sequential decision making under user uncertainty. To verify the correctness of probabilistic action taking at the MBS, we can find out the correlation between each action's selection probability

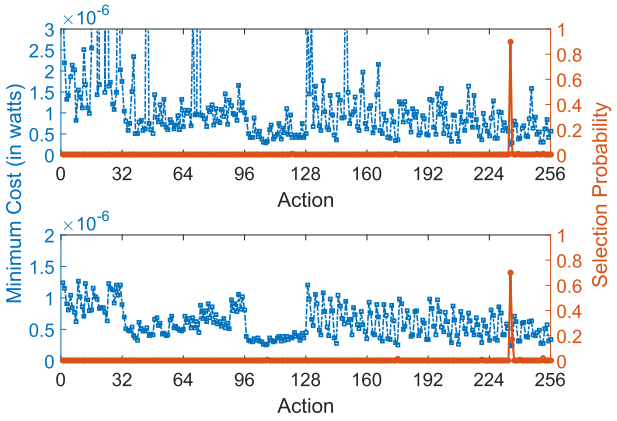


Fig. 7. Strategy correctness of our approach ($t = 800$): (a) $K = 22$ (top); (b) $K = 28$ (bottom).

and its feedback cost. For each possible action $A^j \in \mathcal{A}^J$, we can obtain its selection probability $\pi_t(A^j)$ and its current minimum cost $U_t^*(A^j)$ in a certain stage $t$. The distribution of $\pi_t$ and that of $U_t^*$ can be illustrated together as in Fig. 7. We can see that the high-probability actions match exactly with the low-cost ones given a value of $K$, and the selection probability peaks at the least costly action if $t$ is large enough. The best action stands out more easily for a smaller $K$ in Fig. 7a, since there are more alternative options available for a larger $K$ in Fig. 7b. Note that when $K$ is fairly small, not all the possible actions ensure a feasible solution to our joint optimization problem due to overcrowding in shared spectrum. If an infeasible action is taken, our approach returns a huge cost that exceeds the normal range, as in Fig. 7a, so that this action will no longer be considered in the following stages. Furthermore, the refinement of $\pi_t$ with $t$ is explained in Fig. 8. We can see that a few "good" actions in Fig. 8a are eventually narrowed down to the "best" one in Fig. 8b. This indeed helps our approach converge to a final joint strategy. Therefore, the strategy correctness of our approach can be guaranteed, especially after a sufficient number of stages.

### *C. Power Consumption*

Our approach aims to minimize the weighted sum of all users' power consumption (transmit power levels). To evaluate

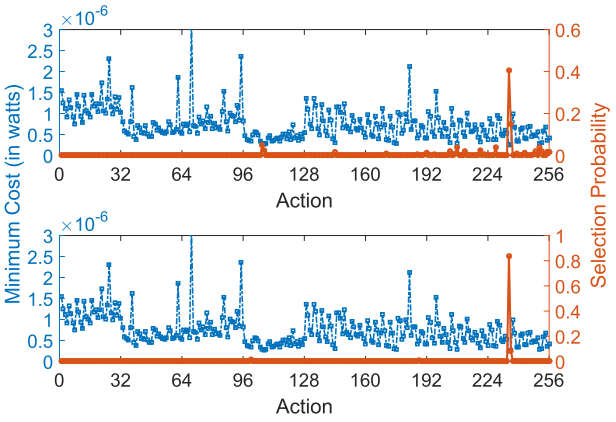


Fig. 8. Impact of $t$ on strategy correctness ($K = 24$): (a) $t = 400$ (top); (b) $t = 800$ (bottom).

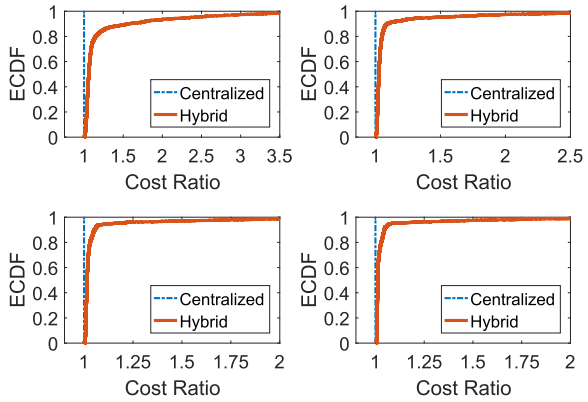


Fig. 9. Power consumption of our approach ($t_0 = 1000$, $\tau = 1500$): (a) $K = 22$ (top-left); (b) $K = 24$ (top-right); (c) $K = 26$ (bottom-left); (d) $K = 28$ (bottom-right).

the gain in transmit power saving, we compare our hybrid multi-agent approach with a fully centralized counterpart. This benchmark approach operates with perfect global knowledge, including hidden user demands, and centrally manages all the issues of SCP, UCA, CRA, TPC. Hence, it derives the optimal solution to our joint optimization problem iteratively, instead of operating in a multi-stage manner. Due to the intractability of the original problem, this centralized approach still follows our logic of problem decoupling. It addresses the two subproblems through a clustering-based revision of Algorithm 1 [16], [34] and a cooperative-game-based revision of Algorithm 2, respectively. After a number of random runs, we can work out the empirical cumulative distribution function (ECDF) of the ratio of the hybrid's objective values $U_t$ for $t = t_0, \ldots, t_0 + \tau$ to the centralized's optimal objective value $U^*$, as shown in Fig. 9. We can see that our hybrid approach achieves nearly as good performance as the centralized counterpart most of the time, even though our approach makes decisions probabilistically. Moreover, our approach allows online decision making without a priori global knowledge and supports computational workload offloading. Furthermore, the refinement of $\pi_t$ for $t = t_0, \ldots, t_0 + \tau$ has its impact on the ECDF of $U_t$, which is shown in Fig. 10. We can see that as our hybrid approach evolves with richer

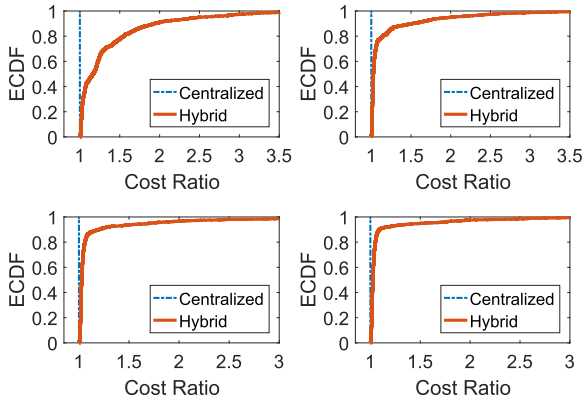


Fig. 10. Impact of $t_0$ on ECDF of power consumption ($K = 24$, $\tau = 300$): (a) $t_0 = 300$ (top-left); (b) $t_0 = 600$ (top-right); (c) $t_0 = 900$ (bottom-left); (d) $t_0 = 1200$ (bottom-right).

TABLE II
SPECTRAL EFFICIENCY OF OUR APPROACH

| | $N = 20$ | $N = 40$ | $N = 60$ | $N = 80$ |
|---|---|---|---|---|
| $M = 5$ | 9.8 | 19.4 | 28.6 | 37.3 |
| $M = 7$ | 9.3 | 18.3 | 27.0 | 35.1 |

user knowledge and wiser action taking, its performance gap with the centralized counterpart becomes narrower. Therefore, our approach can operate efficiently and effectively with a modest loss of optimality in power consumption.

### D. Spectral Efficiency

Our approach minimizes the users' levels of interference power as well as their levels of transmit power, so it is expected to bring a further advantage in spectrum sharing in the uplink. To evaluate the gain in mutual interference mitigation while enabling spectrum reuse among co-located small cells, we can examine the feasibility of accommodating all the users under limited spectrum availability. If $K$ is too small, as mentioned above, it is likely that none of the possible actions achieves a feasible solution to our joint optimization problem. Hence, we can find out the minimum $K$, say $K^*$, that ensures at least one feasible solution. After multiple runs for a combination of $M$ and $N$, the average value of $K^*$ is computed and is presented in Table II. Here we assume that each user takes one channel at a time. We can see that our approach supports the sharing of one channel among more than two users on average, i.e., $\frac{N}{K^*} > 2$, so such spectral efficiency is as good as that in [39]. Generally, a smaller $K^*$ can be achieved for a certain $N$ if given a larger $M$ thanks to greater spectrum reuse. In fact, the mobility of small cells can bring extra benefits to spectrum sharing, such as keeping inter-cell interference distances long enough and making intra-cell transmission distances as short as possible. The deployment of stationary small cells, however, can only be viewed as a feasible solution and cannot be further optimized. Therefore, the spectral efficiency of our approach can be favorable while jointly optimizing small cell deployment and uplink resource allocation.

### *E. Implementation Considerations*

For practical applications, our approach should be implemented in a plug-and-play fashion. As user devices are temporarily gathered outdoors, the MBS's global decisions have to rely on minimum priori user knowledge. Serving as a prerequisite for on-demand deployment of small cells, user demands need to be learnt first and are generalized here for regular uplink transmissions. If data usage patterns of user devices can be obtained in detail, small cell services can be further customized, such as collecting different types of data in different ways. In addition, user locations need to be known to enable on-demand, location-based services. As each small cell moves to cover its target spot, the SBS's local decisions do not have to require timely inter-cell coordination. To support soft frequency reuse in the uplink, user weights can be defined based on transmission/interference distances (instead of propagation gains), and channel quality can be evaluated only based on local interference measurements. Once such system knowledge has been gradually accumulated, our approach can adapt its global and local strategies accordingly.

Even if given perfect global knowledge, a hybrid approach should still be preferred to a completely centralized approach, in consideration of the need for online decision making. Most outdoor hotspots only last for several hours, so the optimal but computationally infeasible solution is not useful. Instead, our approach adopts some heuristics. The complexity of Algorithm 1 is mainly determined by that of centralized optimization for SCP (Lemma 1), which can be approximated as alternating minimization of intra-cell transmission distances. This results in multiple tractable iterations. Given user clusters for UCA, this approximation can be achieved by initially placing each SBS above, e.g., the centroid of its associated user cluster. The complexity of Algorithm 2 is mainly contributed by that of sorting operations for CRA and that of distributed optimization for TPC (Lemma 2). These do not involve sophisticated computations. If there are too many users making the system overly complex, the most demanding ones can be selected to enjoy small cell services while the other ones still receive macro-cell services. As a result, the original problem can be heuristically solved in each stage. The duration of a stage should be short to help capture user dynamics and take prompt actions, while it should not be too short to give sufficient set-up time. Although small cells can hover or remain still to offer reliable outdoor coverage, they may change their hovering points during convergence process. Long-distance movement of a SBS can be avoided by restricting its candidate user clusters for UCA within a certain vicinity. Ideally, each stage can take two or three seconds, which allow small cells to be rapidly deployable in a few minutes and enable a typical UAV to move at least ten meters. In summary, our approach can be implemented in practical situations.

## VIII. Conclusion

In this paper, we focus on spectrum sharing among rapidly deployable small cells in the uplink, which is emphasized for uploading-intensive outdoor hotspots. This requires to deal with a user-centric, online joint optimization of small cell deployment and uplink resource allocation, and requires a low-complexity solution. We have proposed a hybrid multi-agent approach, which operates with two tiers of sequential decision making under user uncertainty. The leading MBS runs Algorithm 1 to derive a mixed strategy for SCP and UCA, based on which the following SBSs run Algorithm 2 independently to construct an equilibrium solution for CRA and TPC. We have proved that the convergence of our approach to a joint strategy is guaranteed for each stage and for a finite time horizon. Our approach is further validated by simulations on the aspects of convergence behavior, strategy correctness, power consumption, and spectral efficiency. We have shown that our hybrid approach achieves nearly as good performance as its centralized counterpart, but our approach is advantageous in that it allows online decision making without a priori global knowledge and supports computational workload offloading.

## References


[1] N. Bhushan *et al.*, "Network densification: The dominant theme for wireless evolution into 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 82–89, Feb. 2014.

[2] A. Gotsis, S. Stefanatos, and A. Alexiou, "UltraDense networks: The new wireless frontier for enabling 5G access," *IEEE Veh. Technol. Mag.*, vol. 11, no. 2, pp. 71–78, Jun. 2016.

[3] "Mass event optimization: Hotspot LTE capacity with evolution to 5G," Nokia, Espoo, Finland, Nokia White Paper C401-012004-WP-201606-1-EN, Jun. 2016.

[4] "5G systems: Enabling the transformation of industry and society," Ericsson, Stockholm, Sweden, Ericsson White Paper UEN 284 23-3251 rev B, Jan. 2017.

[5] Y. Zeng, R. Zhang, and T. J. Lim, "Wireless communications with unmanned aerial vehicles: Opportunities and challenges," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 36–42, May 2016.

[6] S. Hayat, E. Yanmaz, and R. Muzaffar, "Survey on unmanned aerial vehicle networks for civil applications: A communications viewpoint," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 4, pp. 2624–2661, 4th Quart., 2016.

[7] I. Bor-Yaliniz and H. Yanikomeroglu, "The new frontier in RAN heterogeneity: Multi-tier drone-cells," *IEEE Commun. Mag.*, vol. 54, no. 11, pp. 48–55, Nov. 2016.

[8] A. Otto, N. Agatz, J. Campbell, B. Golden, and E. Pesch, "Optimization approaches for civil applications of unmanned aerial vehicles (UAVs) or aerial drones: A survey," *Networks*, vol. 72, no. 4, pp. 411–458, Dec. 2018.

[9] S. Chandrasekharan *et al.*, "Designing and implementing future aerial communication networks," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 26–34, May 2016.

[10] A. Al-Hourani, S. Kandeepan, and S. Lardner, "Optimal LAP altitude for maximum coverage," *IEEE Wireless Commun. Lett.*, vol. 3, no. 6, pp. 569–572, Dec. 2014.

[11] X. Zhou, J. Guo, S. Durrani, and H. Yanikomeroglu, "Uplink coverage performance of an underlay drone cell for temporary events," in *Proc. IEEE ICC Workshops*, May 2018, pp. 1–6.

[12] X. Zhou, S. Durrani, J. Guo, and H. Yanikomeroglu, "Underlay drone cell for temporary events: Impact of drone height and aerial channel environments," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1704–1718, Apr. 2019.

[13] R. I. Bor-Yaliniz, A. El-Keyi, and H. Yanikomeroglu, "Efficient 3-D placement of an aerial base station in next generation cellular networks," in *Proc. IEEE ICC*, May 2016, pp. 1–5.

[14] M. Alzenad, A. El-Keyi, F. Lagum, and H. Yanikomeroglu, "3-D placement of an unmanned aerial vehicle base station (UAV-BS) for energy-efficient maximal coverage," *IEEE Wireless Commun. Lett.*, vol. 6, no. 4, pp. 434–437, Aug. 2017.

[15] M. Alzenad, A. El-Keyi, and H. Yanikomeroglu, "3-D placement of an unmanned aerial vehicle base station for maximum coverage of users with different QoS requirements," *IEEE Wireless Commun. Lett.*, vol. 7, no. 1, pp. 38–41, Feb. 2018.

[16] B. Galkin, J. Kibilda, and L. A. DaSilva, "Deployment of UAV-mounted access points according to spatial user locations in two-tier cellular networks," in *Proc. IFIP/IEEE Wireless Days*, Mar. 2016, pp. 1–6.

[17] J. Lyu, Y. Zeng, R. Zhang, and T. J. Lim, "Placement optimization of UAV-mounted mobile base stations," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 604–607, Mar. 2017.

[18] S. Sharafeddine and R. Islambouli, "On-demand deployment of multiple aerial base stations for traffic offloading and network recovery," *Comput. Netw.*, vol. 156, pp. 52–61, Jun. 2019.

[19] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Drone small cells in the clouds: Design, deployment and performance analysis," in *Proc. IEEE GLOBECOM*, Dec. 2015, pp. 1–6.

[20] E. Kalantari, H. Yanikomeroglu, and A. Yongacoglu, "On the number and 3D placement of drone base stations in wireless cellular networks," in *Proc. IEEE VTC-Fall*, Sep. 2016, pp. 1–6.

[21] C. Zhang and W. Zhang, "Spectrum sharing for drone networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 1, pp. 136–144, Jan. 2017.

[22] A. Kumbhar, I. Güvenç, S. Singh, and A. Tuncer, "Exploiting LTE-advanced HetNets and FeICIC for UAV-assisted public safety communications," *IEEE Access*, vol. 6, pp. 783–796, 2017.

[23] Z. Hu, Z. Zheng, L. Song, T. Wang, and X. Li, "UAV offloading: Spectrum trading contract design for UAV-assisted cellular networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 9, pp. 6093–6107, Sep. 2018.

[24] Y. Zeng and R. Zhang, "Energy-efficient UAV communication with trajectory optimization," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 3747–3760, Jun. 2017.

[25] J. Lyu, Y. Zeng, and R. Zhang, "UAV-aided offloading for cellular hotspot," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 3988–4001, Jun. 2018.

[26] H. He, S. Zhang, Y. Zeng, and R. Zhang, "Joint altitude and beamwidth optimization for UAV-enabled multiuser communications," *IEEE Commun. Lett.*, vol. 22, no. 2, pp. 344–347, Feb. 2018.

[27] S. Jeong, O. Simeone, and J. Kang, "Mobile edge computing via a UAV-mounted cloudlet: Optimization of bit allocation and path planning," *IEEE Trans. Veh. Technol.*, vol. 67, no. 3, pp. 2049–2063, Mar. 2018.

[28] D. Ebrahimi, S. Sharafeddine, P.-H. Ho, and C. Assi, "UAV-aided projection-based compressive data gathering in wireless sensor networks," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1893–1905, Apr. 2019.

[29] M. Samir, S. Sharafeddine, C. Assi, T. M. Nguyen, and A. Ghrayeb, "Trajectory planning and resource allocation of multiple UAVs for data delivery in vehicular networks," *IEEE Netw. Lett.*, vol. 1, no. 3, pp. 107–110, Sep. 2019.

[30] S.-F. Chou, Y.-J. Yu, and A.-C. Pang, "Mobile small cell deployment for service time maximization over next-generation cellular networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 6, pp. 5398–5408, Jun. 2017.

[31] S.-F. Chou, Y.-J. Yu, A.-C. Pang, and T.-A. Lin, "Energy-aware 3D aerial small-cell deployment over next generation cellular networks," in *Proc. IEEE VTC-Spring*, Jun. 2018, pp. 1–5.

[32] A. Fotouhi, M. Ding, and M. Hassan, "DroneCells: Improving 5G spectral efficiency using drone-mounted flying base stations," 2017, *arXiv:1707.02041*. [Online]. Available: https://arxiv.org/abs/1707.02041

[33] Q. Wu, Y. Zeng, and R. Zhang, "Joint trajectory and communication design for multi-UAV enabled wireless networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 2109–2121, Mar. 2018.

[34] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Mobile Internet of Things: Can UAVs provide an energy-efficient mobile architecture?" in *Proc. IEEE GLOBECOM*, Dec. 2016, pp. 1–6.

[35] P. Yang, X. Cao, C. Yin, Z. Xiao, X. Xi, and D. Wu, "Proactive drone-cell deployment: Overload relief for a cellular network under flash crowd traffic," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 10, pp. 2877–2892, Oct. 2017.

[36] Q. Zhang, M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Machine learning for predictive on-demand deployment of UAVs for wireless communications," in *Proc. IEEE GLOBECOM*, Dec. 2018, pp. 1–6.

[37] M. Chen, M. Mozaffari, W. Saad, C. Yin, M. Debbah, and C. S. Hong, "Caching in the sky: Proactive deployment of cache-enabled unmanned aerial vehicles for optimized quality-of-experience," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 5, pp. 1046–1061, May 2017.

[38] A. Morgado, K. M. S. Huq, S. Mumtaz, and J. Rodriguez, "A survey of 5G technologies: Regulatory, standardization and industrial perspectives," *Digit. Commun. Netw.*, vol. 4, no. 2, pp. 87–97, Apr. 2018.

[39] B. Gao, J.-M. J. Park, and Y. Yang, "Uplink soft frequency reuse for self-coexistence of cognitive radio networks," *IEEE Trans. Mobile Comput.*, vol. 13, no. 6, pp. 1366–1378, Jun. 2014.

[40] B. Gao *et al.*, "Incentivizing spectrum sensing in database-driven dynamic spectrum sharing," in *Proc. IEEE INFOCOM*, Apr. 2016, pp. 1–9.

[41] S. Maghsudi and E. Hossain, "Multi-armed bandits with application to 5G small cells," *IEEE Wireless Commun.*, vol. 23, no. 3, pp. 64–73, Jun. 2016.

[42] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The nonstochastic multiarmed bandit problem," *SIAM J. Comput.*, vol. 32, no. 1, pp. 48–77, 2002.

[43] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Found. Trends Mach. Learn.*, vol. 5, no. 1, pp. 1–22, Dec. 2012.

[44] E. Maskin and J. Tirole, "Markov perfect equilibrium: I. Observable actions," *J. Econ. Theory*, vol. 100, no. 2, pp. 191–219, Oct. 2001.

[45] B. Gao, Y. Yang, and J.-M. J. Park, "A credit-token-based spectrum etiquette framework for coexistence of heterogeneous cognitive radio networks," in *Proc. IEEE INFOCOM*, Apr./May 2014, pp. 2715–2723.

**Bo Gao** (M'11) received the Ph.D. degree in computer engineering from Virginia Tech, Blacksburg, VA, USA, in 2014. He was an Assistant Professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, from 2014 to 2017. He was a Visiting Researcher with the School of Computing and Communications, Lancaster University, Lancaster, U.K., from 2018 to 2019. He is currently an Associate Professor with the School of Computer and Information Technology, Beijing Jiaotong University, Beijing. He has directed a number of research projects supported by the National Natural Science Foundation of China (NSFC) or other funding agencies. He is a member of ACM. His research interests include wireless networking, dynamic spectrum sharing, mobile edge computing, and multiagent systems.

**Lingyun Lu** received the Ph.D. degree in computer science from the School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China. She was an Associate Professor with the School of Computer and Information Technology, Beijing Jiaotong University, where she is currently an Associate Professor with the School of Software Engineering. She has led or participated in a number of research projects funded through government or industry sponsors, including the National Natural Science Foundation of China (NSFC). Her current research interests include cross-layer network optimization, heterogeneous networks, and multiuser signal processing.

**Ke Xiong** (M'14) received the B.S. and Ph.D. degrees in computer science from Beijing Jiaotong University, Beijing, China, in 2004 and 2010, respectively. He was a Post-Doctoral Research Fellow with Tsinghua University, Beijing, from 2010 to 2013, and a Visiting Scholar with the University of Maryland, College Park, MD, USA, from 2015 to 2016. He is currently a Professor with the School of Computer and Information Technology, Beijing Jiaotong University. He has published more than 120 academic articles in referred journals and conferences. His current research interests include wireless cooperative networks, wireless powered networks, and network information theory. He has served on the editorial boards for a number of international journals. He is a member of the China Computer Federation (CCF) and a Senior Member of the Chinese Institute of Electronics (CIE).

**Jung-Min (Jerry) Park** (F'17) received the Ph.D. degree in electrical and computer engineering from Purdue University in 2003. He is currently a Professor with the Department of Electrical and Computer Engineering, Virginia Tech and the Site Director of the NSF Industry-University Cooperative Research Center (I-UCRC) called Broadband Wireless Access and Applications Center (BWAC). He served as an Executive Committee Member for the U.S. National Spectrum Consortium (NSC) from 2016 to 2018. NSC is a large consortium of wireless industry stakeholders and universities collaborating with multiple U.S. federal government agencies through a $1.25 billion agreement to support the development of advanced spectrum access technologies. His research interests include dynamic spectrum sharing, emerging wireless technologies, including the IoT and V2X, wireless security and privacy, and applied cryptography. His current or recent research sponsors include the National Science Foundation (NSF), National Institutes of Health (NIH), Defense Advanced Research Projects Agency (DARPA), Army Research Office (ARO), Office of Naval Research (ONR), and several industry sponsors. He is currently serving on the editorial boards for a number of IEEE journals, and is actively involved in the organization of a number of flagship conferences. He was a recipient of the 1998 AT&T Leadership Award, the 2008 NSF Faculty Early Career Development (CAREER) Award, the 2008 Hoeber Excellence in Research Award, the 2014 Virginia Tech College of Engineering Faculty Fellow Award, the 2015 Cisco Faculty Research Award, and the 2017 Virginia Tech College of Engineering Dean's Award for Research Excellence. He is currently serving as the Steering Committee Chair for the IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN). He is an IEEE Fellow for his contributions to dynamic spectrum sharing, cognitive radio networks, and security issues.

**Yaling Yang** (M'06) received the Ph.D. degree in computer science from the University of Illinois at Urbana–Champaign in the summer of 2006. She has concentrated her research on the design, modeling, and analysis of networking systems and security systems. She has been named the Faculty Fellow of the Virginia Tech's College of Engineering in 2016. She is currently a Professor with the Bradley Department of Electrical and Computer Engineering, Virginia Tech. She was a recipient of the NSF Faculty Early Career Award. She has been the principle investigator of nine NSF-funded projects.

**Yuwei Wang** received the M.S. degree in electrical engineering from the Beijing Institute of Technology, Beijing, China, and the Ph.D. degree in computer science from the University of Chinese Academy of Sciences, Beijing. He is currently a Senior Engineer (equivalent to Associate Professor) with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing. His current research interests include next-generation network architecture, mobile edge computing, and cloud computing.