IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 22, 2017, accepted June 25, 2017, date of publication July 5, 2017, date of current version July 31, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2723680

An Energy-Saving Algorithm With Joint User Association, Clustering, and On/Off Strategies in Dense Heterogeneous Networks

LUN TANG, WEILI WANG, YANG WANG, (Member, IEEE), AND QIANBIN CHEN, (Senior Member, IEEE)
Key Laboratory of Mobile Communication, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

Corresponding author: Weili Wang (1961797154@qq.com)

This work was supported by the National Natural Science Foundation of China under Grant 61571073.

ABSTRACT Green networks, which is put forward for the environmental and economic benefits, has received much attention recently because of the vast energy cost in wireless cellular networks. To reduce the energy consumption and simultaneously guarantee the service performance of the dense heterogeneous networks, this paper proposes an energy-saving algorithm with joint user association, clustering, and ON/OFF strategies. First, for the user association subproblem, an optimal association policy, which is related to load balancing and energy efficiency, is designed for the new arriving user equipment (UE) and re-associated UE. Second, based on the locations and load of the base stations (BSs), the clustering subproblem is modeled as an integer linear programming, and the near-optimal clustering results are obtained by using the semi-definite programming. Finally, an intra-cluster ON/OFF strategy for the switching ON/OFF subproblem is designed in which the chosen BSs to be switched OFF are decided by their load effect to other BSs in the clusters. The simulation results demonstrate that, compared with the traditional approaches, the clustering-based energy-saving algorithm can reduce the average network cost by 25.2%–66.7% for different network load conditions.

INDEX TERMS Dense HetNets, energy-saving, switching on/off, clustering, user association.

I. INTRODUCTION

The increasing of UEs and required data rates has triggered vast expansion of network sizes and resulted in significantly increased energy cost [1]–[3]. The rising energy cost has severe environmental and economic effects, which stimulates the research passion for green networks in both academia and industry [4], [5]. Since the energy consumption of the BSs accounts for a significant portion of the whole energy used in cellular networks, nearly 60% $\sim$ 80% [6], [7], the deployment of low-cost and high-capacity small BSs (SBSs) in the coverage of macro BSs (MBSs) has been introduced as a promising solution to reduce the energy consumption of the whole networks while satisfying the quality of service (QoS) requirements of the UEs [8], [9].

User association problem is an important issue in the HetNets that contains both MBSs and SBSs for their massive disparities in cell sizes [10]. Suppose a relatively uniform UE distribution, the unequal cell sizes will result in very uneven loads in a classical max-SINR user association. To optimize the aforementioned problem, there have been many studies [10]–[15] on user association policies for the HetNets, and they basically considered two objectives as their metrics for selecting the serving BSs, (i) load balancing [10]–[13] and (ii) energy efficiency [14], [15]. It should be noted that the user association strategy developed in the paper, which considers both load balancing and energy efficiency, is an extended form of [10].

The BSs in the wireless networks are usually deployed to satisfy the QoS required by the UEs in peak periods, while the user traffic is unevenly distributed in both time and space [16]. The relevant research has pointed out that the average time portion is more than 40% of one day when the traffic is below 10% of the peak, and the underutilization of BSs results in the large waste of bandwidth and energy resource [17]. Therefore, switching off some underutilized BSs has substantial potential to reduce the energy consumption. Different algorithms have been designed to achieve energy conservation [18]–[21]. In [18], the BS switching operations were formulated as a Markov decision process, and the transfer actor-critic method was adopted to decide the on/off state of the BSs. In [19], two greedy algorithms were proposed to minimize the number of active BSs under the limitation of the UEs' minimal rate requirements. In [20], the total energy consumption of the BSs was minimized in two steps by taking the switching energy cost into consideration. For the delay will increase inevitably when some BSs in the networks are in the off state, delay-constrained energy-optimal BS sleeping policies were designed in [21] based on a queuing theoretic model. However, existing works rarely dealt with the joint optimization of delay and energy in dense HetNet environments.

In the dense HetNets, the interference among the BSs is serious and the implemented complexity of the switching on/off strategy is also high. By reference to the existing researches, it can be found that grouping the dense deployed BSs into clusters is an important way to lower the complexity of the algorithm while eliminating the mutual interference effectively [22]–[24]. Hence, the paper considers introducing the cluster patterns into the proposed energy-saving algorithm.

Based on the lightly loaded downlink transmission of dense HetNets, a clustering-based energy-saving algorithm is proposed in the paper, and the key contributions of the paper can be summarized as follows:

(1)Firstly, a theoretical framework is developed for green networks that considers user association and on/off strategies jointly. Specifically, a total cost minimization problem under the load and average outage probability constraints is formulated to strive a tradeoff between the energy consumption and average delay. Through adjusting a tradeoff parameter, the framework can adapt to different network environments.

(2)Secondly, to alleviate the interference among BSs and lower the complexity of on/off strategies, the network clustering is introduced to the optimization model. The paper decomposes the last model into three subproblems, (i) user association, (ii) network clustering and (iii) on/off strategies, and solves these problems one by one. Firstly, the paper designs an optimal user association strategy for the new arriving UEs, by which the joint minimization of the energy consumption and average delay can be achieved. Then, a distance-load based clustering algorithm is proposed in the paper. The algorithm can group the two BSs into the same cluster for their close distance to effectively eliminate the mutual interference by using orthogonal resource allocation, and the two BSs, whose gap of the load is large can also be partitioned into the same cluster to facilitate the BSs being switched off and saving energy. Finally, the paper proposes an intra-cluster on/off strategy that the selected BSs to be switched off are decided by their load effects to other BSs in the clusters.

(3)Thirdly, the paper evaluates the performance of the proposed clustering-based energy-saving algorithm with extensive simulations, and compares it with three other approaches, which are the conventional network operation, the random on/off algorithm and the similar energy-saving algorithm to the proposed one but without clusters. The results demonstrate that the proposed algorithm can significantly reduce the energy consumption of the networks while guaranteeing the QoS requirements of UEs.

The rest of the paper is organized as follows. The system model and problem formulation are presented in Section II. Section III provides the optimal user association strategy for user association subproblem. In Section IV, the load-distance based clustering subproblem is solved with the SDP. Then, the intra-cluster switching on/off subproblem is discussed in Section V. Section VI provides the simulation results and the conclusions and future work are drawn in Section VII.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. SYSTEM MODEL

The paper considers a dense heterogeneous cellular network whose set of BSs, denoted by B, includes MBSs and SBSs. The focus is on downlink communication, i.e., from BSs to UEs. Assume that the BSs are uniformly distributed over a two-dimensional network region $\boldsymbol{A}$ and x represents a location coordinate in the region. Let A_b denote the coverage area of BS b, then given any UE, if its coordinate vector satisfies $x \in A_b$, and it can be obtained that the associated BS of the UE is BS b.

The UEs in the network region have various QoS requirements for their different date rates. Assume that $\gamma(x)$ denotes the traffic arrival rate of any UE located in x, and $B_{\text{on}} \subset B$ is the set of active BSs. Then, let $s_b(x)$ be the transmission rate of any UE at location x, served by BS b, $b \in B_{\text{on}}$. Therefore, the traffic load density at location x can be defined as $\upsilon_b(x) = \gamma(x)/s_b(x)$, and the load on active BS b is equal to the integral of load density in the coverage area of itself:

$$\Gamma(B_{\text{on}}) = \left\{ \boldsymbol{\rho} | \rho_b = \int_{A_b} \frac{\gamma(x)}{s_b(x)} m_b(x)\mathrm{d}x \text{ , } \forall b \in B_{\text{on}} \right\}, \quad (1)$$

where $\boldsymbol{\rho} = (\rho_1, \rho_2, \ldots, \rho_{|B|})$, $m_b(x) \in \{0, 1\}$ is an indicator function defined as $m_b(x) = 1$ if any UE at location x is associated with BS b, $m_b(x) = 0$, otherwise. In the light of [10], Γ (B_{on}) can be known as a convex set.

In the paper, the load ρ_b of BS b represents the total transmission time from itself to all UEs covered by it. Furthermore, the average traffic flow of BS b can be expressed as $\rho_b/(1 - \rho_b)$, which is proportional to the average delay of BS b [10].

The effective transmission power of BS b is proportional to its load ρ_b, and when it uses a transmission power of P_b, the effective transmission power is ${P_b}^{\text{Work}} = \rho_b P_b$. Besides the effective transmission power ${P_b}^{\text{Work}}$, an active BS b consumes additional power P_b^{Base} to operate its radio frequency components and baseband unit [25]. From an energy-saving perspective, some BSs might be in off state, and suppose that the BSs have zero energy consumption during the off state, then the total power consumption model of BSs set B can be defined as:

$$P_b^{\text{Total}} = \begin{cases} 0, & if\ b \in B \backslash B_{\text{on}}, \\ \dfrac{1}{\chi_b} \rho_b P_b + P_b^{\text{Base}}, & if\ b \in B_{\text{on}}, \end{cases} \quad (2)$$

where χ_b is used to represent the losses in the power source and cooling units and the efficiency of the power amplifier of BS b.

The focus is on the downlink transmission, and the channels between UEs and BSs are modeled as additive Gaussian white noise (AGWN) channels with noise variance N_0. Based on Shannon's formula, the achievable data rate of any UE at location $x \in A_b$, served by BS b, is defined as $s_b(x) = W\log_2(1 + \mathrm{SINR}_b(x))$ when given bandwidth W, and the $\mathrm{SINR}_b(x)$ is given by:

$$\mathrm{SINR}_b(x) = \frac{P_b g_b(x)}{\sum\limits_{a \in B_{\mathrm{on}} \backslash b} P_a{}^{\mathrm{Work}} g_a(x) + N_0}, \quad b \in B_{\mathrm{on}}, \tag{3}$$

where $g_b(x)$ denotes the channel gain between BS b and UE at location x.

B. PROBLEM FORMULATION

For energy-saving purpose, some BSs in the network region might be switched off, and the service performance of active BSs will inevitably deteriorate. Therefore, the objective is to minimize the energy consumption of the network while reducing the average delay of all BSs during the downlink transmission. According to section *A*, it can be known that the power consumption of BS b consists of two parts: the effective transmission power $P_b{}^{\mathrm{Work}}$ which is proportional to its load ρ_b and the fixed power consumption P_b^{Base}. For the off-state BSs have zero energy consumption, the cost function of energy can be given by:

$$\psi(\boldsymbol{\rho}, B_{\mathrm{on}}) = \sum_{b \in B_{\mathrm{on}}} \left(\frac{1}{\chi_b} \rho_b P_b + P_b^{\mathrm{Base}} \right). \tag{4}$$

To achieve the objective of reducing the energy consumption while ensuring the service performance of the BSs, the delay-optimal criterion is introduced to guarantee the service performance of the BSs. The cost function of delay-optimal performance is given by :

$$\phi(\boldsymbol{\rho}, B_{\mathrm{on}}) = \sum_{b \in B_{\mathrm{on}}} \frac{\rho_b}{1 - \rho_b}. \tag{5}$$

By reference to the Little's law, minimizing the cost function of delay-optimal performance is equivalent to minimize the average delay of all BSs [18].

The reduction of energy consumption can be achieved by switching off some BSs in the network region, which will cause that the UEs covered by them associate with other active BSs. Since the re-associated BSs are no longer the optimal ones, the average outage probability is introduced to avoid too poor QoS of the UEs previously covered by the inactive BSs. For any BS to be switched off, if the average outage probability of the UEs associated with it is higher than a certain threshold, the switching-off request will be rejected.

The outage probability [26], [27] of any UE at location x is given by:

$$\begin{aligned} P^{\mathrm{out}}(x) &= \mathrm{Prob}(\mathrm{SINR}_a(x) \le \mathrm{SINR}^{\mathrm{th}}) \\ &= 1 - e^{-\lambda_a(x) N_0} \prod_{k \in B_{\mathrm{on}} \backslash a} \frac{\lambda_k(x)}{\lambda_k(x) + \mathrm{SINR}^{\mathrm{th}} \lambda_a(x)}, \\ &\qquad\qquad x \in A_b, \end{aligned} \tag{6}$$

where $\lambda_a(x) = 1/g_a(x)P_a$, $a \in B_{\mathrm{on}}$, and a is the suboptimal associated BS for any UE at location x.

Then, the average outage probability of UEs associated with BS b, which is to be switched off is given by:

$$P_{\text{ave - b}}^{\mathrm{out}} = \frac{1}{U_b} \int_{A_b} P^{\mathrm{out}}(x) dx, \quad b \in B \backslash B_{\mathrm{on}}, \tag{7}$$

where U_b is the total number of UEs who are associated with BS b.

In the paper, the problem is to find the optimal set of active BSs (B_{on}) and the corresponding optimal vector of user association ($\boldsymbol{\rho}$), which can minimize the energy consumption of the network while ensuring the service performance of the BSs. Meanwhile, the average outage probability is introduced to ensure the overall QoS of the UEs previously covered by the inactive BSs. Guaranteeing the load of active BSs to not exceed a certain threshold is also a necessary limitation to make the networks stable. Thus, the minimization of the network cost, which is equal to the weighted sum of power consumption and average delay of all active BSs, is given by the following optimization problem:

$$\begin{aligned} &\min_{\boldsymbol{\rho}, B_{\mathrm{on}}} \sum_{b \in B_{\mathrm{on}}} \left\{ \left(\eta \frac{\rho_b}{1 - \rho_b} \right) + \left(\frac{1}{\chi_b} \rho_b P_b + P_b^{\mathrm{Base}} \right) \right\} \\ &s.t.\ C1 : P_{\text{ave - b}}^{\mathrm{out}} \le P^{\mathrm{th}}, \quad \forall b \in B \backslash B_{\mathrm{on}}, \\ &\qquad C2 : \rho_{\mathrm{b}} \le \rho^{\mathrm{th}}, \quad \forall b \in B_{\mathrm{on}}, \end{aligned} \tag{8}$$

where η is a tradeoff parameter that can be dynamically adjusted to control the impact of energy consumption and average delay on the network cost. Hence, through changing the value of η, the above formulation can adapt to different network environments.

The above optimization objective is to minimize the overall network cost in the limitation of the load of active BSs and the average outage probability of re-associated UEs. Here, constraint (C1) ensures that the average outage probability of the re-associated UEs is less than P^{th} ($P^{\mathrm{th}} < 1$), such that the QoS of re-associated UEs can be satisfied approximately; Constraint (C2) guarantees the load of active BSs is not more than ρ^{th} ($\rho^{\mathrm{th}} < 1$) to make the networks stable.

The complexity of switching on/off strategies is $2^{|B|}$ [19]. With the increase number of BSs in the networks, the complexity of this problem will increase exponentially. Even if the number of BSs is only moderate-sized, solving the optimization problem will be very complicated. In dense HetNets, the BS amount is high and the interference among the BSs is also serious. Therefore, grouping all BSs into the set of clusters is considered, and in each cluster, orthogonal

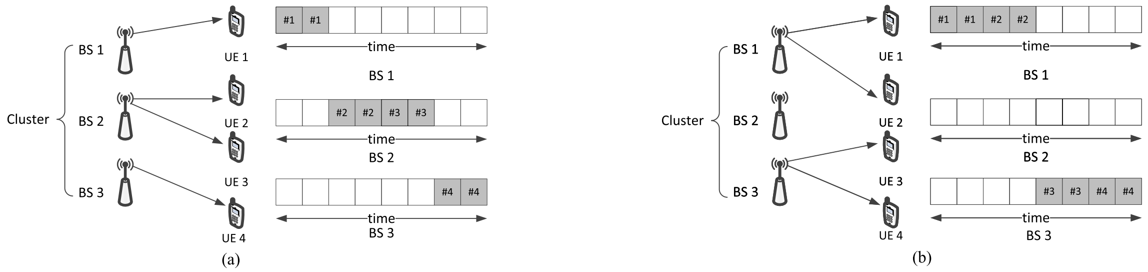


FIGURE 1. Intra-cluster orthogonal resource allocation to avoid interference. The users can be offloaded in each cluster to facilitate saving energy. (a) When all the BSs are in on state. (b) When the BS 2 is in the off state.

resources are allocated among the BSs to eliminate the mutual interference effectively as illustrated in Fig. 1. Furthermore, the switching on/off strategies are carried out within each cluster to lower the complexity of the optimization problem. After the clusters are grouped, the paper assumes that all re-associated UEs choose the suboptimal BSs within their own clusters. Given the set of clusters $\bar{\boldsymbol{S}} = \{S_1, S_2, \ldots, S_{|\bar{\boldsymbol{S}}|}\}$, then the outage probability of the UEs covered by the BSs to be turned off is given by:

$$\begin{aligned} P^{\text{out}}(x) &= \text{Prob}(\text{SNR}_a \le \text{SNR}^{\text{th}}) \\ &= \{1 - \exp[-\lambda_a(x) \cdot \text{SNR}^{\text{th}} N_0]\}, \\ &\qquad\qquad x \in A_b, \quad a \in S_i^{\text{on}}, \qquad (9) \end{aligned}$$

where a is the intra-cluster suboptimal associated BSs for any UE in location x, and S_i^{on} is the set of active BSs in cluster S_i, which includes BS b.

The average outage probability is given by:

$$P^{\text{out}}_{\text{ave - b}} = \frac{1}{U_b} \int_{A_b} P^{\text{out}}(x) dx, \quad b \in S_i \backslash S_i^{\text{on}}. \qquad (10)$$

Given the grouped clusters, the optimization goal is expressed as the following problem:

$$\begin{aligned} &\min_{\boldsymbol{\rho}, B_{\text{on}}, \bar{\boldsymbol{S}}} \sum_{S_i \in |\bar{\boldsymbol{S}}|} \sum_{b \in S_i} \left\{ \left(\eta \frac{\rho_b}{1-\rho_b} \right) + \left(\frac{1}{\chi_b} \rho_b P_b + P_b^{\text{Base}} \right) \right\} \\ &s.t. \; C1: \; |S_i| \ge 1, \quad \forall S_i \in \bar{\boldsymbol{S}}, \\ &\quad C2: S_i \cap S_j = \emptyset, \quad \forall S_i, S_j \in \bar{\boldsymbol{S}}, \; S_i \ne S_j, \\ &\quad C3: P^{\text{out}}_{\text{ave - b}} \le P^{\text{th}}, \quad \forall b \in S_i \backslash S_i^{\text{on}}, \; S_i \in \bar{\boldsymbol{S}}, \\ &\quad C4: \rho_b \le \rho^{\text{th}}, \quad \forall b \in S_i^{\text{on}}, \; S_i \in \bar{\boldsymbol{S}}. \qquad (11) \end{aligned}$$

Here, constraints (C1) and (C2) are introduced for clustering that ensure any BS is a member of only one cluster. Constraint (C3) guarantees the QoS of re-associated UEs. Finally, to make the networks stable, constraint (C4) limits the maximum load for the active BSs.

In the paper, the downlink transmission consists of three phases: (1) all the UEs determine the active BSs with which they wish to associate. (2) The on/off state selections of all active BSs are carried out in each cluster, and all re-associated UEs determine the BSs to re-associate with in their own clusters. (3) Re-cluster and perform on/off state selections in each cluster until the network stability is reached.

III. USER ASSOCIATION

Since the time-scale for user association is much smaller than that of determining the set of active BSs, and there is no direct relationship between new UEs' association and clustering results, the user association problem can be solved for any given set of active BSs B_{on}. When the set of active BSs is given, the fixed power consumption term $\sum_{b \in B_{\text{on}}} P_b^{\text{Base}}$ is a constant value. Although the capacity of BS is limited, the assumption that the QoS of new UEs can always be satisfied is reasonable for the included load constraint, i.e., there is no limit to the outage probability for new UEs. Thus, the user association subproblem is equivalent to the following objective:

$$\begin{aligned} &\min_{\boldsymbol{\rho}} \sum_{b \in B_{\text{on}}} \left\{ \left(\eta \frac{\rho_b}{1-\rho_b} \right) + \left(\frac{1}{\chi_b} \rho_b P_b \right) \right\} \\ &s.t. \; \rho_b \le \rho^{\text{th}}, \quad \forall b \in B_{\text{on}}. \qquad (12) \end{aligned}$$

Let $\rho^* = (\rho_1^*, \rho_2^*, \ldots, \rho_{|B_{\text{on}}|}^*)$ denote the optimal load vector which satisfies the optimization problem (12), and further denote the optimal user association at location x by $i^*(x)$. Then, the following objective is to get the solution to the problem (12), i.e., developing the solution to the optimal association strategy $i^*(x)$.

Theorem 1: If the optimal load vector $\rho^* = (\rho_1^*, \rho_2^*, \ldots, \rho_{|B_{\text{on}}|}^*)$ satisfying the problem (12) exists, then the optimal association strategy of any UE at location x is:

$$i^*(x) = \arg\max_{b \in B_{\text{on}}} \frac{s_b(x)}{\eta (1-\rho_b^*)^{-2} + (1/\chi_b) P_b}, \quad \forall x \in \boldsymbol{A}. \qquad (13)$$

Its implication is as follows. When $\eta = 0$, the user association is determined by the energy efficiency of the network. However, as $\eta \to \infty$, the UEs associate with the BSs that can better satisfy their QoS requirements.

The *Theorem* 1 is proved in part *B*.

A. USER ASSOCIATION POLICY

The user association policy involves two parts.

User Equipment: At the start of the kth period, UEs receive the broadcast control messages $\rho^{(k)}$ from BSs. Then, a new service request for a UE located at x determines the associated BS utilizing the following rule given by:

$$i(x) = \arg\max_{b\in B_{\text{on}}} \frac{s_b(x)}{\eta(1-\rho_b^{(k)})^{-2} + (1/\chi_b)P_b}, \quad \forall x \in \boldsymbol{A}. \tag{14}$$

This shapes a new spatial partition $\boldsymbol{A}^{(k)} = \{A_1^{(k)}, A_2^{(k)}, \ldots, A_{|B_{\text{on}}|}^{(k)}\}$, where $A_b^{(k)}$ defines the coverage area of BS b at the kth period. In particular, $A_b^{(k)}$ depends on the broadcast load $\rho^{(k)}$ as follows:

$$A_b^{(k)} = \left\{ x \in \boldsymbol{A} | b = \arg\max_{j\in B_{on}} \frac{s_j(x)}{\eta(1-\rho_j^{(k)})^2 + (1/\chi_j)P_j} \right\}. \tag{15}$$

Base Station: During the kth iteration period, BSs calculate their average load:

$$T_b(\rho^{(k)}) = \min[\int_{A_b^{(k)}} \upsilon_b(x)dx, \rho^{\text{th}}], \quad b \in B_{\text{on}}, \tag{16}$$

where $T(\rho) = (T_1(\rho), T_2(\rho), \ldots, T_{|B_{\text{on}}|}(\rho))$ is a continuous mapping defined on $[0, \rho^{\text{th}}]$ to itself. Note that the mapping $T(\rho)$ is a mathematical model denoting user association dynamics at any period. Specifically, when BSs broadcast the load ρ, and the user association policy is followed, finally, the BSs will measure a new load vector $T(\rho)$.

B. THE OPTIMALITY OF USER ASSOCIATION POLICY

Note that if $\rho^{(k)}$ converges, it must converge to a fixed point, which is a solution satisfying $\rho^* = T(\rho^*)$ [10].

In the following, the paper will prove that $T(\cdot)$ has a unique fixed point ρ^* corresponding to the solution to the optimization problem (12), i.e., the optimal load vector satisfying it.

Proof: For $T(\rho)$ is defined on $[0, \rho^{\text{th}}]$ and a continuous mapping to itself, referring to Brouwer's fixed point theorem, a solution to $T(\rho^*) = \rho^*$ must exist.

In the following, that ρ^* is the optimal solution for the user association problem (12) will be proved.

Since Γ (B_{on}) is a convex set and the objective function of (12) is a convex function, i.e., problem (12) is a convex function over a convex set, if ρ^* satisfies the in equation:

$$\langle \nabla\phi(\rho^*), \Delta\rho^* \rangle \geq 0, \tag{17}$$

then ρ^* is the optimal solution to problem (12), and $\Delta\rho^* = \rho - \rho^*$ for any $\rho \in \Gamma$ (B_{on}).

Let $m(x)$ and $m^*(x)$ be the indicator functions for ρ and ρ^* respectively. Based on the association rule, $m^*(x)$ is given by the following formula:

$$m_b^*(x) = \mathbf{1}\left\{ b = \arg\max_{j\in B_{\text{on}}} \frac{s_j(x)}{\eta(1-\rho_j^*)^{-2} + (1/\chi_j)P_j} \right\}, \tag{18}$$

and the inner product (17) can be computed as:

$$\begin{aligned} &\langle \nabla\phi(\rho^*), \Delta\rho^* \rangle \\ &= \sum_{b\in B_{\text{on}}} [\eta \frac{1}{(1-\rho_b^*)^2} + \frac{1}{\chi_b}P_b)](\rho_i - \rho_i^*) \\ &= \int_{\boldsymbol{A}} \gamma(x) \sum_{b\in B_{\text{on}}} \{ \frac{[\eta(1-\rho_b^*)^{-2} + (1/\chi_b)P_b)]}{s_b(x)} \\ &\qquad \cdot (m_b(x) - m_b^*(x))\} dx, \end{aligned} \tag{19}$$

for $m^*(x)$ is an indicator for the maximization of $\frac{s_b(x)}{\eta(1-\rho_b^*)^{-2}+(1/\chi_b)P_b}$, we can deduce that:

$$\begin{aligned} &\sum_{b\in B_{\text{on}}} \frac{\eta(1-\rho_b^*)^{-2} + (1/\chi_b)P_b}{s_b(x)} m_b(x) \\ &\quad \geq \sum_{b\in B_{\text{on}}} \frac{\eta(1-\rho_b^*)^{-2} + (1/\chi_b)P_b}{s_b(x)} m_b^*(x). \end{aligned} \tag{20}$$

In conclusion, the optimal association rule, which is put forward in *Theorem* 1, can meet the inner product condition (17), i.e., ρ^* is the optimal solution to problem (12).

IV. CLUSTER FORMATION

To solve the optimization objective (11), an available clustering algorithm and an effective energy-saving algorithm within each cluster are indispensable. For grouping all BSs into clusters, the network is mapped to a weighted graph $G = \{\boldsymbol{B}, \boldsymbol{E}\}$, where $\boldsymbol{B}$ is the set of vertices and $(i, j) \in \boldsymbol{E}$ is the set of edges between two vertices. Each edge (i, j) is allocated with two weights $w_{i,j}^+$ and $w_{i,j}^-$, which represent the degree of similarity and difference between two vertices, respectively. The target is to find a partition that any two BSs are grouped into the same cluster for high similarity and small difference.

In the paper, the optimization objective is to minimize the weighted sum of energy consumption and average delay. According to the formula (4), it can be seen that the size of energy consumption is proportional to the number of active BSs, and referring to formulas (1) and (5), the average delay of the BSs are closely related with their service rates. Therefore, the two BSs, whose gap of load is large, are expected to partition into one cluster for facilitating the BSs being switched off and saving energy. Meanwhile, the two BSs, whose distance is close, are expected to be in the same cluster, and in each cluster, the orthogonal resource allocation is used to eliminate the interference and reduce the average delay. Therefore, the paper sets $w_{i,j}^+ = |\rho_i - \rho_j|$ to represent the motive why two BSs would be in the same cluster, and $w_{i,j}^- = d_{i,j}$ to represent the resistance between them.

Based on these, the clustering problem is given by the following optimization objective:

$$\begin{aligned}
&\max_{x_{i,j}} \sum_{i\in B_{\text{on}}} \sum_{j\in B_{\text{on}}} \vartheta|\rho_i-\rho_j|x_{i,j}+d_{i,j}(1-x_{i,j})\\
&s.t.\ C1:\ x_{i,i}=1, \quad \forall i\in B_{\text{on}},\\
&\quad C2:\ x_{i,j}=x_{j,i}, \quad \forall i,j\in B_{\text{on}},\\
&\quad C3:\ x_{i,j}+x_{j,k}-x_{i,k}\le 1, \quad \forall i,j,k\in B_{\text{on}}:\\
&\qquad\qquad k>i, j\ne i,k,\\
&\quad C4:\ \sum_{j\in B_{\text{on}}} x_{i,j}\le M, \quad \forall i\in B_{\text{on}},\\
&\quad C5: x_{i,j}\in\{0,1\}, \quad \forall i,j\in B_{\text{on}},
\end{aligned} \tag{21}$$

where ϑ is a tradeoff parameter balancing the impact of $w^{+}_{i,j}$ and $w^{-}_{i,j}$ on the clustering results. When $\vartheta = 0$, the partition only depends on the distance among the BSs, since the distance is the penalty for putting two BSs into the same cluster, the resulting partition is to put one BS in a single cluster, i.e., a cluster contains only one BS. When $\vartheta \to \infty$, the load gap among BSs determines the clustering results, since the load gap is the motive for forming clusters, if the constraint (C4) is ignored, all BSs will be grouped into the same cluster, i.e., the whole network is the one and only cluster.

In the optimization problem (21), constraint (C1) indicates that a BS can only be the member of one cluster. Constraint (C2) is the symmetry condition, which means that if one cluster contains i but not j, then the cluster containing j must not contain i. Constraint (C3) is a triangular condition, which specifies that if i , j are in the same cluster and j, k are in the same cluster, then i and k must also be in the same cluster. Constraint (C4) is a ceiling on the cluster size, and in the paper, the maximum cluster size can not exceed M. Finally, in constraint (C5), $x_{i,j}$ is a binary clustering variable, $x_{i,j} = 1$ when i and j are in the same cluster, and $x_{i,j} = 0$, otherwise.

The optimization objective (21) is an ILP, which is NP-hard. Its optimal solution can be obtained using Brach and Bound (BnB). Nevertheless, BnB has an exponential complexity, apparently it is impractical in dense HetNets. Therefore, another approach depending on SDP is employed to solve the problem. SDP has only one additional semi-definite constraint compared to general linear programming.

Based on the SDP, the clustering problem can be expressed as the following formulation:

$$\begin{aligned}
&\max_{x_{i,j}} \sum_{i\in B_{\text{on}}} \sum_{j\in B_{\text{on}}} \vartheta|\rho_i-\rho_j|x_{i,j}+d_{i,j}(1-x_{i,j})\\
&s.t.\ C1:\ x_{i,i}=1, \quad \forall i\in B_{\text{on}},\\
&\quad C2:\ x_{i,j}+x_{j,k}-x_{i,k}\le 1, \quad \forall i,j,k\in B_{\text{on}}:\\
&\qquad\qquad k>i, j\ne i,k,\\
&\quad C3:\ \sum_{j\in B_{\text{on}}} x_{i,j}\le M, \quad \forall i\in B_{\text{on}},\\
&\quad C4: x_{i,j}\ge 0, \quad \forall i,j\in B_{\text{on}},\\
&\quad C5: \mathbf{X}=(x_{i,j})\succeq 0, \quad \forall i,j\in B_{\text{on}},
\end{aligned} \tag{22}$$

where $\mathbf{X} = \{x_{i,j}\}, i \in B_{\text{on}}, j \in B_{\text{on}}$ is the clustering matrix.

In such a problem, constraint (C4) indicates that the binary variable $x_{i,j}$ has been relaxed and constraint (C5) ensures that the elements in clustering matrix meet the semi-definite constraint, i.e., the matrix must be symmetric and each element contained in the matrix is greater than or equal to 0.

The symmetric matrix $\mathbf{X}$ can be written as $\mathbf{X} = \mathbf{V}^T\mathbf{V} = \mathbf{QD}^{1/2}(\mathbf{QD}^{1/2})^T$, where $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_{|B_{\text{on}}|}]$, $\mathbf{Q}$ is the matrix whose each column is a eigen vector of $\mathbf{X}$, and $\mathbf{D}$ is the diagonal matrix whose diagonal elements are the eigen values of $\mathbf{X}$. CVX, a software package for formulating and solving the convex problems [28] is applied to solve the optimization objective (22). It is worth mentioning that the eigen decomposition for the matrix $\mathbf{X}$, which is solved by the toolboxes, has a complexity of $O(|B_{\text{on}}|^3)$. For the size limitation to each cluster in clustering, the matrix $\mathbf{X}$ is sparse when the number of active BSs is large, and the sparsity in the matrix $\mathbf{X}$ can decrease the complexity significantly and improve the efficiency of the computation.

After obtaining the clustering matrix $\mathbf{X}$, it will be transformed into the vector format $\mathbf{X} = \mathbf{V}^T\mathbf{V}$. For the element values in matrix $\mathbf{X}$ are not binary numbers and fall within the interval (0, 1), the Randomized Rounding (RR) [29] is employed to obtain the final clustering results. The RR solution can be achieved by generating L unit random vectors $\mathbf{r}_i^T = (r_{i1}, r_{i2}, \ldots, r_{i|B_{\text{on}}|})$, $i = 1, 2, \ldots, L$, where each element value of $\mathbf{r}_i$ is randomly fetched from $N(0, 1)$, the Gaussian distribution with mean 0 and variance 1. For clustering accurately, the number of random vectors L must satisfy the inequation of $2^L \ge B_{\text{on}}$. The L vectors can give rise to 2^L clusters at most, $c_1, c_2, \cdots, c_{2^L}$. Then, the mapping of BSs into clusters is done as follows:

$$\begin{aligned}
c_1 &= \{i\in B_{\text{on}}: \mathbf{r}_1\mathbf{v}_i\ge 0, \cdots, \mathbf{r}_{L-1}\mathbf{v}_i\ge 0, \mathbf{r}_L\mathbf{v}_i\ge 0\}\\
c_2 &= \{i\in B_{\text{on}}: \mathbf{r}_1\mathbf{v}_i\ge 0, \cdots, \mathbf{r}_{L-1}\mathbf{v}_i\ge 0, \mathbf{r}_L\mathbf{v}_i< 0\}\\
c_3 &= \{i\in B_{\text{on}}: \mathbf{r}_1\mathbf{v}_i\ge 0, \cdots, \mathbf{r}_{L-1}\mathbf{v}_i< 0, \mathbf{r}_L\mathbf{v}_i\ge 0\}\\
&\cdots\\
c_{2^L} &= \{i\in B_{\text{on}}: \mathbf{r}_1\mathbf{v}_i< 0, \cdots, \mathbf{r}_{L-1}\mathbf{v}_i< 0, \quad \mathbf{r}_L\mathbf{v}_i< 0\}.
\end{aligned} \tag{23}$$

Let U_{optimal} and U_{RR} denote the optimal function value of (21) and (22) respectively, then, an α-approximation algorithm, where U_{RR} satisfies $U_{\text{RR}} \ge \alpha U_{\text{optimal}}$ ($\alpha < 1$) can be obtained by performing RR technique. There are more chances of U_{RR} getting very close to the optimal value U_{optimal} when the entire procedure is repeated for *iteration* number of times, according to Algorithm 1.

V. INTRA-CLUSTER SWITCHING ON/OFF MECHANISM

Given the formed clusters $\bar{\mathbf{S}} = \{S_1, S_2, \ldots, S_{|\bar{\mathbf{S}}|}\}$, the next goal is to solve the problem of the on/off state selections of BSs in each cluster. The optimization objective for switching

Algorithm 1 SDP-Based Clustering Algorithm With Multiple Iterations

```
Input: $\vartheta$, $|\rho_i - \rho_j|, i, j \in B_{\text{on}}$, $d_{i,j}, i, j \in B_{\text{on}}$
Output: The optimal clustering results $\mathbf{X}_{\text{optimal}}$
Solve the relaxed optimization problem using software package, CVX, and obtaining $\mathbf{X}$
for i=1:iteration do
    Generate an $L$ number of the random vectors $\mathbf{r}_1, \mathbf{r}_2, \cdots, \mathbf{r}_L$
    Discard the zero clusters in $c_1, c_2, \cdots, c_{2L}$, and obtain the initial clustering results $\mathbf{X}_{\text{init}}$
    Bring $\mathbf{X}_{\text{init}}$ into the optimization goal of problem (21), and get the objective function value
end for
Determine an optimal clustering solution $\mathbf{X}_{\text{optimal}}$ that has the largest objective function value
```

on/off is given as follows:

$$\min_{B_{\text{on}},\boldsymbol{\rho}} \sum_{S_i\in\bar{\mathbf{S}}} \sum_{b\in S_i^{\text{on}}} \left(\varphi(\rho_b) + P_b^{\text{Base}}\right)$$

$$s.t.\ P_{\text{ave - b}}^{\text{out}} \le P^{\text{th}}, \quad \forall b \in S_i \backslash S_i^{on},\ S_i \in \bar{\mathbf{S}},$$

$$\rho_{\text{b}} \le \rho^{\text{th}}, \quad \forall b \in S_i^{on},\ S_i \in \bar{\mathbf{S}}, \tag{24}$$

where $\varphi\ (\rho_b) = \eta \dfrac{\rho_b}{1-\rho_b} + \dfrac{1}{\chi_b}\rho_b P_b$.

Theorem 2: the problem (24) is NP-hard.

Proof: Assume that $\mathbf{U}$ is the set of UEs, and $\mathbf{U}_b(\mathbf{U}_b \subset \mathbf{U})$ is the set of UEs covered by BS b. Let $\mathbf{S} = \{\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_{|B|}\}$ represents the set of UE subsets. Then, consider the following simplified problem: suppose that the capacity of a BS $b(\forall b \subset B)$ is unlimited and its cost is $\varphi\ (\rho_b) + P_b^{\text{Base}}$, and the objective is to find a set cover $\mathbf{S}^* \subseteq \mathbf{S}$ which satisfies $\bigcup_{\mathbf{U}_b\subset\mathbf{S}^*} \mathbf{U}_b = \mathbf{U}$ and the cost of $\mathbf{S}^*$ is minimized. Then, this problem is equivalent to a weighted set cover problem, which is known to be NP-hard. Since the assumed problem is a subproblem of the optimization objective (24), the problem (24) is NP-hard as well.

In general, obtaining the optimal solution to the NP-hard problem is very difficult, therefore, an effective approximation algorithm is required to get the satisfying quasi-optimal solution to the on/off state selections of BSs.

In the following, an effective intra-cluster switching on/off mechanism is proposed for energy conservation of the networks. The goal is to find a set of active BSs under the limiting conditions of load and average outage probability. When the number of active BSs in each cluster is constant, updating the partition until the network is stable. In cluster S_i, the switching off of a BS b must be carried out within the range where the total cost is always reduced, i.e., the following condition must be satisfied:

$$\sum_{m\in\{S_i-b\}} \left(\varphi_m(\rho_{b\to m} + \rho_m) + P_m^{\text{Base}}\right) \le \sum_{m\in S_i} \left(\varphi(\rho_m) + P_m^{\text{Base}}\right), \tag{25}$$

where $\rho_{b\to m}$ is the load increment of BS m for the switched off of its neighboring BS b.

The left side of the above inequality represents the total cost of the cluster S_i after switching off one of its member b, and the right side shows the total cost of the cluster before switching off any BS. Only when the condition is satisfied that the total cost after turning off one BS is lower than that before, can the operation be called worthwhile, i.e., the switching off of any BS must be carried out within the range where the total cost is always reduced. Since the load ρ is the only variable in the cost function, the aforementioned range for one cluster S_i can be expressed as the load range, which should be satisfied by the other BSs after turning off one BS b in the cluster. Therefore, only when the limiting condition of the load range is satisfied by the active BSs, can the operation of switching off any BS b in the cluster be called worthwhile. In the following, the paper will provide the lower limit of load that satisfies the limiting condition (25) and consider it as the load threshold ρ^{th}.

The cost function is convex, which means that as the load increases, the growth rate of the cost value also increases. Assume that n is an active BS in the cluster S_i, then, the following formula must be held:

$$\begin{aligned}
&\sum_{m\in\{S_i-b\}} \left(\varphi_m(\rho_{b\to m} + \rho_m) + P_m^{\text{Base}}\right) \\
&\quad \le \varphi_n(\rho_b + \rho_n) + P_n^{\text{Base}} + \sum_{m\in\{S_i-b-n\}} \left(\varphi(\rho_m) + P_m^{\text{Base}}\right) \\
&\quad \le \sum_{m\in S_i} \left(\varphi(\rho_m) + P_m^{\text{Base}}\right).
\end{aligned} \tag{26}$$

Simplify the inequation, and the following formula can be obtained:

$$\varphi_n(\rho_b + \rho_n) - (\varphi(\rho_b) + \varphi(\rho_n)) \le P_b^{\text{Base}}, \tag{27}$$

observe the above formula, for $\varphi\ (\rho)$ is a convex function, if $\rho = \rho_b + \rho_n$ is a fixed value, it can be known that the left side of the inequality reaches its maximum if and only if $\rho_b = \rho_n$, that is:

$$\begin{aligned}
&\varphi_n(\rho_b + \rho_n) - (\varphi(\rho_b) + \varphi(\rho_n)) \\
&\quad \le \varphi_n(\rho) - (\varphi_b(\frac{\rho}{2}) + \varphi_n(\frac{\rho}{2})) \\
&\quad = \frac{\eta\rho}{1-\rho} + \frac{\rho P_n}{2\chi_n} - (\frac{2\eta\rho}{2-\rho} + \frac{\rho P_b}{2\chi_b}) \\
&\quad \le P_b^{\text{Base}}.
\end{aligned} \tag{28}$$

The expression of ρ^{th} is given by:

$$\frac{\eta\rho^{\text{th}}}{1-\rho^{\text{th}}} + \frac{\rho^{\text{th}} P_n}{2\chi_n} - (\frac{2\eta\rho^{\text{th}}}{2-\rho^{\text{th}}} + \frac{\rho^{\text{th}} P_b}{2\chi_b}) = P_b^{\text{Base}}. \tag{29}$$

Hence, the operation of switching off BS b in the cluster S_i is limited by the following feasibility constraint:

$$\int_{A_m} \frac{\gamma(x)}{s_m(x)} dx + \int_{A_{b\to m}} \frac{\gamma(x)}{s_m(x)} dx \le \rho^{\text{th}}, \quad m \in S_i^{\text{on}} \backslash b, \tag{30}$$

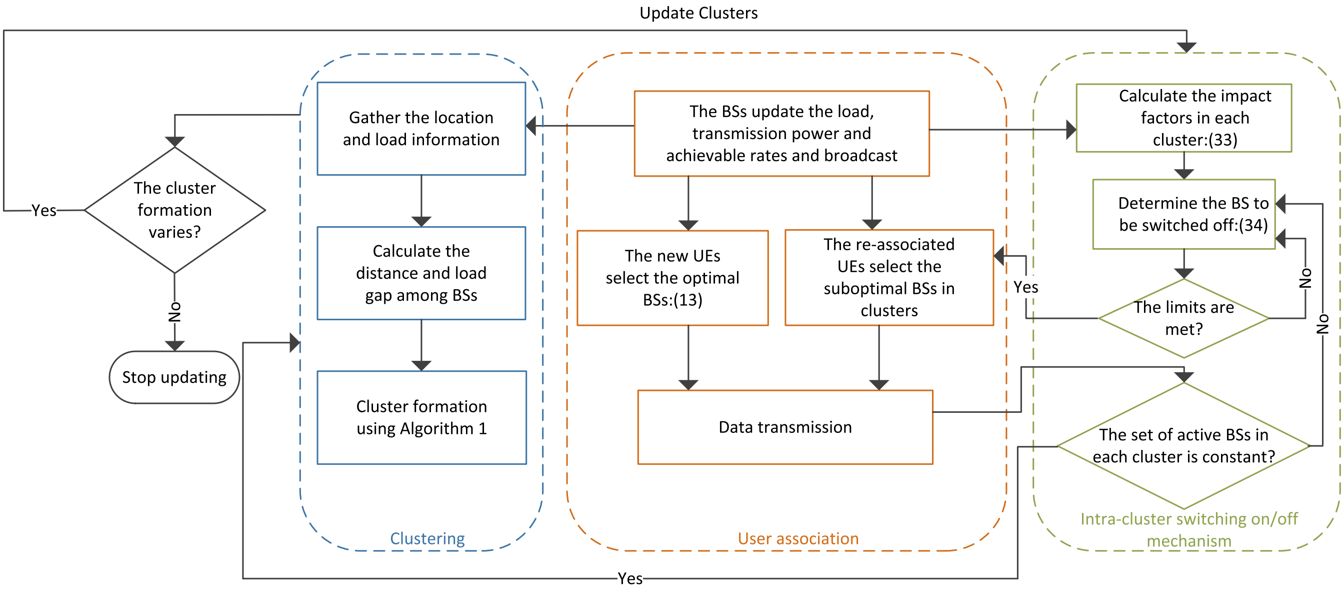

Update Clusters
Yes
The cluster formation varies?
No
Stop updating
Gather the location and load information
Calculate the distance and load gap among BSs
Cluster formation using Algorithm 1
Clustering
The BSs update the load, transmission power and achievable rates and broadcast
The new UEs select the optimal BSs:(13)
The re-associated UEs select the suboptimal BSs in clusters
Data transmission
User association
Calculate the impact factors in each cluster:(33)
Determine the BS to be switched off:(34)
The limits are met?
Yes
No
The set of active BSs in each cluster is constant?
No
Intra-cluster switching on/off mechanism
Yes

FIGURE 2. **The entire flow diagram of the proposed intra-cluster energy-saving algorithm.**

where the BS m can be interpreted as the BS with which the UEs will be re-associated after switching off BS b, and $A_{b\to m}$ is the coverage area of UEs who will be transferred from BS b to the suboptimal BS m when the BS b is turned off.

In order to determine the BSs set that can be switched off in the cluster S_i, the concept of impact factor [30] is introduced, on which the intra-cluster switching on/off strategy is based. The definition of impact factor is given by the following expression:

$$F_b = \max_{m\in S_i^{on}\backslash b} (\rho_m + \rho_{b\to m}), \quad \forall b \in S_i^{\text{on}}. \tag{31}$$

The impact factor of BS b indicates the maximum load after its turning off within the cluster.

The value of $\rho_{b\to m}$ is very hard to be predicted precisely, hence the paper proposes a kind of approximately estimating method. Assume that the additional load of BS m, which transferred from the switching off BS b, is closely related with the original load of BS b and the Euclidean distance $d_{bm}(d_{bm} = ||x_b - x_m||)$ between BS b and m. Furthermore, the additional load of BS m is inversely proportional to d_{bm}, thus the relation can be given by:

$$\rho_{b\to m} = k\rho_b\mu_{bm}, \tag{32}$$

where $k = 1/\sum_{i\in S_i}\mu_{bi}$, $\mu_{bm} = 1/d_{bm}$. The computational complexity of $\rho_{b\to m}$ is relatively low for the locations of BSs can be determined easily. The complete expression of the impact factor for BSs can be given by:

$$F_b = \max_{m\in S_i^{\text{on}}\backslash b} (\rho_m + k\rho_b\mu_{bm}), \quad \forall b \in S_i. \tag{33}$$

The switching on/off strategy is implemented within each cluster, and during each iteration the BS with the minimum value of impact factor is selected to identify if the load and average outage probability limitations can be satisfied by it, i.e., the BS to be switched off should meets the following formula:

$$b^* = \arg\min_{b\in S_i^{\text{on}}} F_b. \tag{34}$$

The next step is to verify whether the selected BS satisfies the load constraint (31) and the average outage probability constraint. If the both constraints can be satisfied, the selected BS can be turned off, and the UEs covered by it choose the suboptimal associated BSs within the cluster through the previously proposed association policy. In each cluster, if all existing active BSs cannot be switched off, then re-cluster and perform the intra-cluster switching on/off mechanism until the network is stable. The entire flow diagram of the proposed intra-cluster energy-saving algorithm is illustrated in Fig. 2. Algorithm 2 shows the detailed steps of the intra-cluster switching on/off mechanism.

VI. SIMULATION RESULTS

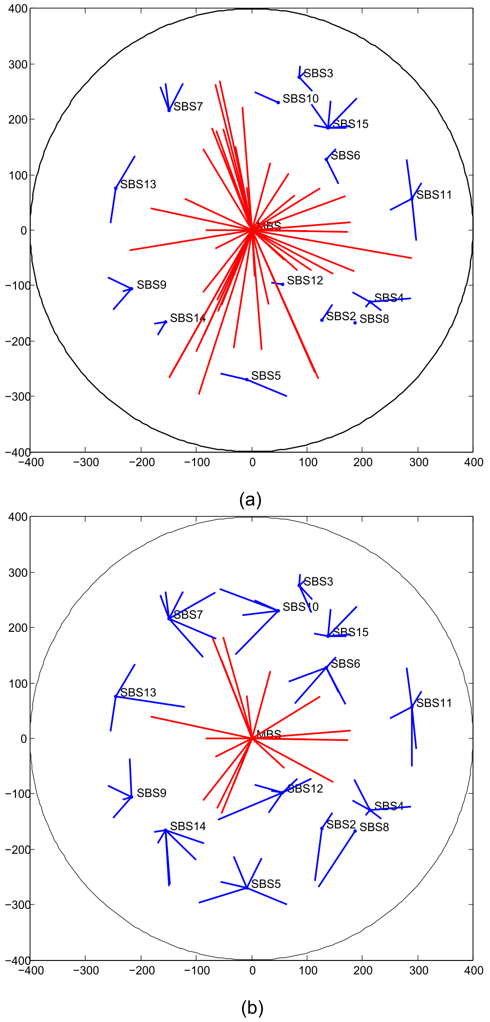
In the numerical simulations, the availability of the three sub-algorithms, which are the optimal energy-delay association, load-distance based clustering and intra-cluster switching on/off algorithms, are verified through extensive simulation analysis. In addition, the joint clustering-based energy-saving algorithm, which consists of the aforementioned three sub-algorithms, is proved to be effective through comparative simulations. A dense HetNet consisting of a single MBS underlaid with a set of SBSs and UEs uniformly distributed over the coverage area is considered for the simulations. The presented results are averaged multiple independent runs with various practical configurations. The parameters used for the simulation results are summarized in Table 1.

Algorithm 2 The Intra-Cluster Switching On/Off Algorithm

```
Input: Clustering results $\bar{\mathbf{S}} = \{S_1, S_2, \ldots, S_{|\bar{\mathbf{S}}|}\}$, load $\rho_b, b \in S_i$, transferred load $\rho_{b\to m}, m \in S_i^{\mathrm{on}}$, load threshold $\rho^{\mathrm{th}}$, average outage probability threshold $P^{\mathrm{th}}$
Output: Stable sets of active BSs
for $i = 1 : |\bar{\mathbf{S}}|$ do
    for $b = 1 : |S_i|$ do
        Select a BS to be switching off: $b^* = \arg\min_{b\in S_i^{\mathrm{on}}} F_b$
        Determine whether or not the two limiting conditions can be met:
        Load constraint:
        $\int_{A_m} \frac{\gamma\ (x)}{s_m(x)}dx + \int_{A_{b\to m}} \frac{\gamma\ (x)}{s_m(x)}dx \le \rho^{\mathrm{th}}, m \in S_i^{\mathrm{on}}\backslash b$
        Outage probability constraint:
        $P^{\mathrm{out}}_{\mathrm{ave\ -\ b}} \le P^{\mathrm{th}}, \forall b \in S_i\backslash S_i^{on}$
        If the both constraints are met, the BS $b^*$ can be turned off, otherwise, continue keeping it active
    end for
end for
Re-cluster the set of active BSs based on Algorithm 1, and continue executing the above loop until the network is stable
```

TABLE 1. Simulation parameters.

Parameter	Value
Carrier frequency,System bandwidth	2GHz,10MHz
Thermal noise(N_0)	-174dBm/Hz
Mean traffic rate($\gamma(x)$)	960kbps
Maximum transmission powers:MBS,SBS	46,30dBm
Efficiency of power unit(χ):MBS,SBS	59%,13.5%
Minimum distances and path loss models(d in km)	
MBS-SBS,MBS-UE	75m,35m
SBS-SBS,SBS-UE	40m,10m
MBS-UE path loss	$128.1 + 37.6\ \log_{10}(d)$
SBS-UE path loss	$140.7 + 36.7\ \log_{10}(d)$



(a)

(b)

FIGURE 3. Resulting association patterns of max-SINR vs. the proposed association algorithm of energy-delay tradeoff. (a) Max-SINR association. (b) Optimal energy-delay association.

Fig. 3 shows two different association results, which are respectively based on the traditional max-SINR association strategy and the optimal energy-delay association strategy proposed in the paper. By comparing the association patterns in Fig. 3a and Fig. 3b, it can be found that the SINR-based approach will associate the majority of UEs with MBS, while many of the SBSs serve very few UEs. The results are caused by the big transmission power difference between MBS and SBS. Therefore, employing SINR-based association approach in dense HetNets will inevitably increase the sojourn time of most UEs associated with MBS and also cause the waste of energy and spectrum resources. The optimal energy-delay association strategy can drive the UEs to choose the associated BSs whose energy consumption and average delay are both acceptable. Therefore, it can be easily analyzed that the optimal association strategy proposed in the paper can reduce the energy consumption of the networks while ensuring the QoS of UEs.

Fig. 4 shows the relationship between the load-distance tradeoff parameter ϑ and the cluster size / the number of clusters in the networks. Through analyzing the clustering results, when the value of ϑ is small, the number of BSs contained by each cluster is also small, and the number of grouped clusters is large correspondingly. It is because the distance, which represents the penalty for clustering, decides the formation of clusters with a small value of ϑ. When the value of ϑ is large, the clustering results is mainly determined by the load difference between any two BSs, which is known as the motive to the partition of the network, therefore, the network is partitioned into few clusters and the number of BSs in each cluster is large.

Fig. 5 shows the converging process of the switching on/off algorithm for different number of UEs as the iteration times increase. When no BS can be switched off in each cluster with current clustering results, re-clustering and the iteration times increase by one. It can be observed that the proposed algorithm converges rapidly within several iteration times. Fig. 5a and Fig. 5b illustrate the convergence of average energy consumption and average delay per BS, respectively. Observing Fig. 5a, it can be found that the average energy

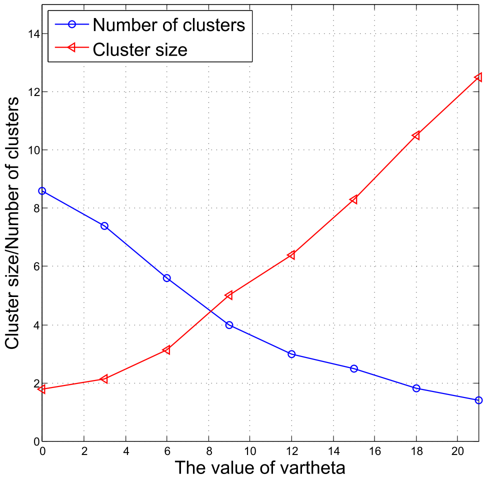


FIGURE 4. Impact of ϑ on the cluster size / number of clusters.

consumption per BS will decrease gradually for the increased number of off-state BSs as the switching on/off algorithm performs iteratively. Conversely, Fig. 5b indicates that the average delay of the on-state BSs will increase for the increased number of UEs that each BS serves as the number of off-state BSs increases. Furthermore, it can be found that the simulation scenarios with large number of UEs need few iteration times. It is because that when the number of UEs increases, the number of BSs that can be switched off will decrease, and the iteration times for convergence of energy-saving algorithm will decrease as well.

For verify the effectiveness of the proposed clustering-based energy-saving algorithm, the paper compares it with the conventional network operation, which will be called as "classical algorithm" hereinafter, in which the BSs always transmit. For further comparisons, the paper considers the other two algorithms, a random on/off switching algorithm in which each BS has the equal probability to be switched off and a similar energy saving algorithm to the proposed approach but without forming clusters. In the following, these two approaches and the proposed algorithm are referred to as "random on/off", "on/off without clustering" and "on/off with clustering", respectively.

Fig. 6 shows the comparisons among the different algorithms about the average cost per BS and the number of outage users when the total number of UEs varies. Observing Fig. 6a, it can be found that as the number of UEs increases, the average energy consumption and average delay per BS increase, and the average cost increases as well. It is an additional gain to find that the proposed on/off with clustering algorithm has an approximate linear energy-saving results with the number of UEs, which manifests that the proposed algorithm performs steadily as the number of UEs varies. Furthermore, the simulation results show that the on/off with clustering algorithm can largely reduce the average cost through striking a tradeoff between the energy consumption and the average delay. As the number of UEs varies, compared to the classical algorithm, the reduction ranges

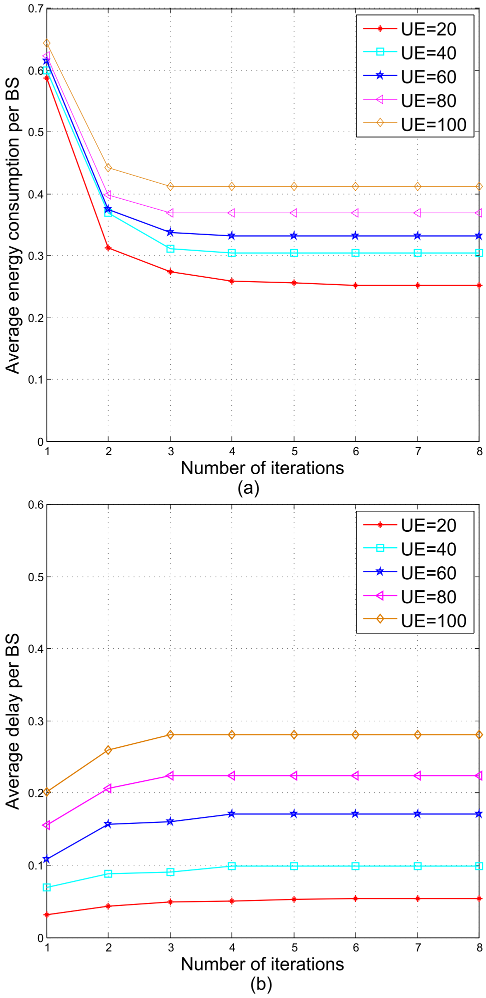


FIGURE 5. The converging process of the switching on/off algorithm as the iteration times increase. (a) The convergence of energy consumption. (b) The convergence of average delay.

of average cost in random on/off, on/off without clustering, and on/off with clustering approaches are 4.8%∼35.5%, 6.1%∼37.5%, and 25.2%∼66.7%, respectively. Therefore, it can be confirmed that the on/off with clustering algorithm can achieve a large reduction in average cost.

Fig. 6b indicates that the three energy-saving algorithms will all slightly decrease the QoS of UEs when compared with the classical algorithm, in which the BSs always transmit. The on/off without clustering algorithm has the smaller number of outage users compared with the random on/off algorithm and the on/off with clustering algorithm for its average outage probability limitation and less off-state BSs, respectively. The on/off with clustering algorithm has the similar number of outage users with the Random on/off algorithm despite of more energy conservation than it. Fig. 6b further prove the effectiveness of the clustering-based energy-saving algorithm proposed in the paper.

Fig. 7 reflects the changing trends of average energy consumption and average delay per BS as the energy-delay tradeoff parameter η varies from 100∼800. When the number of

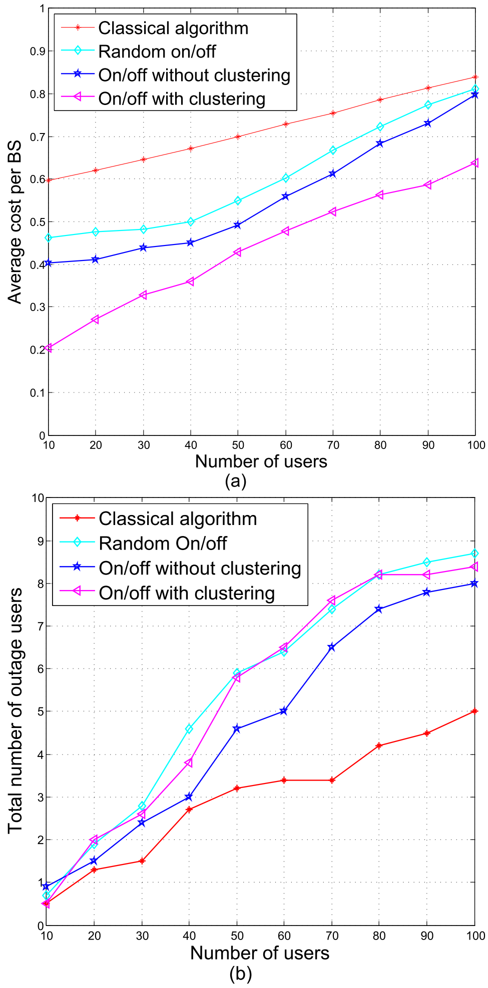


FIGURE 6. Comparison among the four algorithms about the average cost per BS and number of outage users. (a) Average cost per BS as the number of UEs varies. (b) Total number of outage UEs as the number of UEs varies.

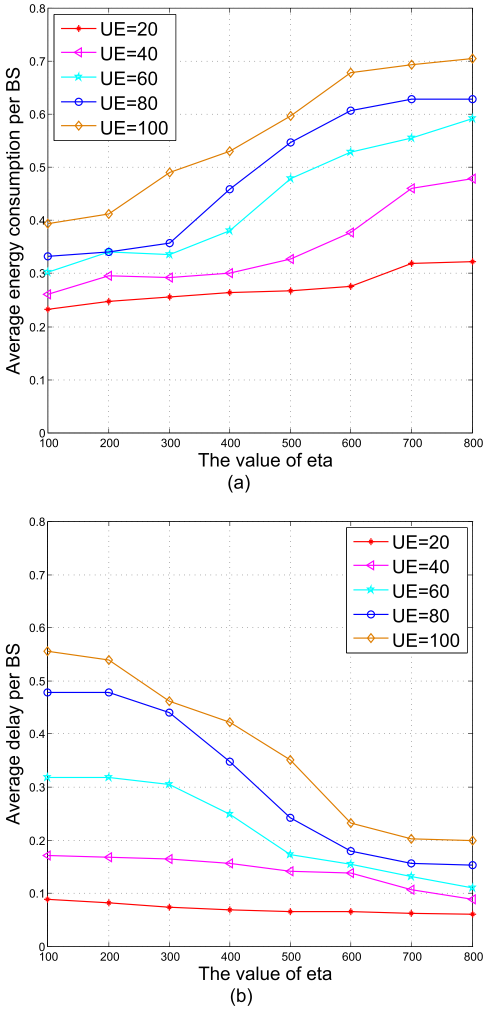


FIGURE 7. Relationship between η and average energy consumption/average delay per BS. (a) The impact of η on the average energy consumption per BS. (b) The impact of η on the average delay per BS.

users increase, the effect of η on the average energy consumption and average delay per BS increase accordingly. It can be seen that the effect is little when the number of users is 20. Comparing Fig. 7a and Fig. 7b, with the increase of η, the average delay per BS will reduce while the average energy consumption per BS will increase, and both will eventually stabilize. This is because the average delay occupies a dominant position in the average cost of BSs when the value of η is large, and the effect of energy consumption is relatively weakened. Therefore, the size of η can be dynamically adjusted to satisfy the different QoS requirements for various network environments according to their sensitivity to delay and energy.

The energy-saving algorithm proposed in the paper is solved through being decomposed into three sub-problems: user association, clustering and on/off strategies for BSs. The joint algorithm is not superfluous in simulating the practical networks for the internal relationship among the three subproblems. The efficient user association criterion is indispensable to the energy-saving algorithm, because the users, which were associated with the off-state BSs previously, require a rule to re-associated with the on-state BSs while ensuring their QoS. In dense heterogeneous networks, interference is a non-negligible problem, and the introduction of cluster can eliminate the mutual interference through orthogonal resource allocation in each cluster while lowering the complexity of on/off strategies. Therefore, an available algorithm should take some relevant problems into joint consideration in simulating the practical networks.

VII. CONCLUSION AND FUTURE WORK

A novel clustering-based energy-saving algorithm has been proposed in the paper to minimize the average cost for the downlink transmission of dense HetNets. Firstly, for the user association problem, an optimal association strategy, which is based on the tradeoff between energy and delay, was proposed for the new arriving UEs and re-associated UEs, and that the UEs select the associated BSs through using the strategy could achieve the minimization of energy consumption while ensuring the service performance of the networks. Then, based on the load and distance information, the paper formulated the clustering problem as an integer linear programming, and the near-optimal clustering results were obtained by using the semi-definite programming. Finally, the intra-cluster switching on/off algorithm was executed according to the impact factor of BSs. The paper deduced the lower limit of the load for the switching on/off mechanism, and the network cost would decrease if the load of on-state BSs was larger than the lower limit values after switching off the certain BSs. The introduction of clusters in the proposed algorithm not only could effectively eliminate the mutual interference between adjacent BSs, and also could drop the complexity of the switching on/off algorithm. The simulation results showed that the clustering-based energy-saving algorithm could significantly reduce the average cost of the networks compared with other algorithms.

For the stochastic nature of traffic arriving process in practical networks, the traffic distribution after switching off some BSs is uncertain, and the bursty traffic near the off-state BSs would increase the risk of unstable networks. Therefore, the future research will focus on the traffic forecast before performing the on/off strategies, furthermore, the switching energy cost, which is caused by the on/off state switching of BSs, will also be taken into consideration for its not trivial value.

REFERENCES

[1] M. A. Marsan and M. Meo, "Network sharing and its energy benefits: A study of European mobile network operators," in *Proc. IEEE GLOBECOM*, Dec. 2013, pp. 2561–2567.

[2] Z. Niu, "ANGO: Traffic-aware network planning and green operation," *IEEE Wireless Commun.*, vol. 18, no. 5, pp. 25–29, Oct. 2011.

[3] J. B. Rao and A. O. Fapojuwo, "A survey of energy efficient resource management techniques for multicell cellular networks," *IEEE Commun. Surveys Tut.*, vol. 16, no. 1, pp. 154–180, 1st Quart., 2014.

[4] Y. Chen, S. Zhang, and G. Li, "Fundamental tradeoffs on green wireless networks," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 30–37, Jun. 2011.

[5] Z. Hasan, H. Boostanimehr, and V. K. Bhargava, "Green cellular networks: A survey, some research issues and challenges," *IEEE Commun. Surveys Tut.*, vol. 13, no. 4, pp. 524–540, 4th Quart., 2011.

[6] M. A. Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo, "Optimal energy savings in cellular access networks," in *Proc. 1st Int. Workshop Green Commun.*, Jun. 2009, pp. 1–5.

[7] E. Oh, B. Krishnamachari, X. Liu, and Z. Niu, "Toward dynamic energy efficient operation of cellular network infrastructure," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 56–61, Jun. 2011.

[8] Q. Wang and J. Zheng, "A distributed base station on/off control mechanism for energy efficiency of small cell networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2015, pp. 3317–3322.

[9] P. C. Lin, L. F. G. Casanova, and Y. C. Lin, "Analytical framework for power saving evaluation in two-tier heterogeneous mobile networks," *Wireless Netw.*, vol. 23, no. 4, pp. 985–999, May 2017.

[10] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam, "Distributed *alpha*-optimal user association and cell load balancing in wireless networks," *IEEE/ACM Trans. Netw.*, vol. 20, no. 1, pp. 177–190, Feb. 2012.

[11] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, Jun. 2013.

[12] Z. Mlika, M. Goonewardena, and W. Ajib, "User-base station association in HetSNets: Complexity and efficient algorithms," *IEEE Trans. Veh. Technol.*, vol. 66, no. 2, pp. 1484–1495, Feb. 2017.

[13] Y. Chen, J. Li, and Z. Li, "User association with unequal user priorities in heterogeneous cellular networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 9, pp. 7374–7388, Sep. 2016.

[14] H. Pervaiz, L. Musavian, and Q. Ni, "Joint user association and energy-efficient resource allocation with minimum-rate constraints in two-tier HetNets," in *Proc. IEEE Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2013, pp. 1634–1639.

[15] T. Q. Zhou, Y. M. Huang, and L. X. Yang, "Energy-efficient user association in downlink heterogeneous cellular networks," *IET Commun.*, vol. 10, no. 13, pp. 1553–1561, Aug. 2016.

[16] J. Kim, P.-Y. Kong, N.-O. Song, J.-K. K. Rhee, and S. Al-Araji, "MDP based dynamic base station management for power conservation in self-organizing networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2014, pp. 2561–2567.

[17] J. Zheng, Y. Cai, X. Chen, R. Li, and H. Zhang, "Optimal base stations sleeping in green cellular networks: A distributed cooperative framework based on game theory," *IEEE Trans. Wireless Commun.*, vol. 14, no. 8, pp. 4391–4406, Aug. 2015.

[18] R. Li, Z. Zhao, X. Chen, J. Palicot, and H. Zhang, "TACT: A transfer actor-critic learning framework for energy saving in cellular radio access networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 4, pp. 2000–2011, Apr. 2014.

[19] Y. Yang, L. Chen, W. Dong, and W. Wang, "Active base station set optimization for minimal energy consumption in green cellular networks," *IEEE Trans. Veh. Technol.*, vol. 64, no. 11, pp. 5340–5349, Nov. 2015.

[20] N. Yu, Y. Miao, and L. Mu, "Minimizing energy cost by dynamic switching ON/OFF base stations in cellular networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7457–7469, Nov. 2016.

[21] X. Guo, Z. Niu, and S. Zhou, "Delay-constrained energy-optimal base station sleeping control," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1073–1085, May 2016.

[22] A. Abdelnasser, E. Hossain, and I. K. Dong, "Clustering and resource allocation for dense femtocells in a two-tier cellular OFDMA network," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1628–1641, Mar. 2014.

[23] L. Zhou, X. Hu, and E. Ngai, "A dynamic graph-based scheduling and interference coordination approach in heterogeneous cellular networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 5, pp. 3735–3748, May 2016.

[24] L. Liang, W. Wang, and Y. Jia, "A cluster-based energy-efficient resource management scheme for ultra-dense networks," *IEEE Access*, vol. 65, pp. 6823–6832, Nov. 2016.

[25] S. Samarakoon, M. Bennis, and W. Saad, "Dynamic clustering and ON/OFF strategies for wireless small cell networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 2164–2178, Mar. 2016.

[26] S. Kandukuri and S. Boyd, "Optimal power control in interference-limited fading wireless channels with outage-probability specifications," *IEEE Trans. Wireless Commun.*, vol. 1, no. 1, pp. 46–55, Jan. 2002.

[27] G. Su, B. Chen, and X. Lin, "A submodular optimization framework for outage-aware cell association in heterogeneous cellular networks," *Math. Problems Eng.*, vol. 2016, pp. 1–11, Feb. 2016.

[28] M. Grant and S. Boyd. (Nov. 2015). *CVX: MATLAB Software for Disciplined Convex Programming, Version 3.0 Beta*. [Online]. Available: http://cvxr.com/cvx/beta/

[29] V. V. Vazirani, *Approximation Algorithms*. Berlin, Germany: Springer-Verlag, 2003.

[30] 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.

LUN TANG received the Ph.D. degree in communication and information systems from Chongqing University Posts and Telecommunications, Chongqing, China. He is currently a Professor with the School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications. His research interests include 5G cellular networks, interference management, and small cell networks.

WEILI WANG received the B.Sc. degree from Chongqing University of Posts and Telecommunications, Chongqing, China, in 2015, where she is currently pursuing the M.S. degree in communication and information systems. Her research interests are in 5G cellular networks, with an emphasis on resource allocation techniques for energy efficiency in dense heterogeneous networks.

YANG WANG (M'15) received the M.S. and Ph.D. degrees from The University of Sheffield, U.K., in 2011 and 2015, respectively. He joined the School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications in 2015. His research interests include antennas and propagation, metamaterial, radar signature management, phase modulating microwave structures, and communications networks.

QIANBIN CHEN (M'03–SM'14) received the Ph.D. degree in communication and information systems from the University of Electronic Science and Technology of China, Chengdu, China, in 2002. He is currently a Professor with the School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, and the Director of the Chongqing Key Laboratory of Mobile Communication Technology. He has authored or co-authored more than 100 papers in journals and peer-reviewed conference proceedings, and has co-authored seven books. He holds 47 granted national patents.