

Received September 28, 2018, accepted November 2, 2018, date of publication November 5, 2018, date of current version December 3, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2879740

Optimal Frequency Reuse in HetNets With In-Band Relays

SHENGDA JIN^{1,2,3}, ZHAOWEI ZHU^{1,2,3}, YUECHEN WU³, SADIQ ALI⁴,
AND XILIANG LUO³, (Senior Member, IEEE)

¹Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China

²University of Chinese Academy of Sciences, Beijing 100049, China

³School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China

⁴Department of Electrical Engineering, University of Engineering and Technology, Peshawar 25000, Pakistan

Corresponding author: Xiliang Luo (luoxl@shanghaitech.edu.cn)

This work was supported through the startup fund from ShanghaiTech University under Grant F-0203-14-008.

ABSTRACT In-band relay nodes can be utilized to enhance the coverage of heterogeneous networks in a cost-effective way. However, the wireless backhaul links of in-band relay nodes also consume the limited spectrum resources. Appropriate spectrum resource management/cooperation is necessary to ensure the balanced resource usages between the macro base stations and the relay nodes. In this paper, we derive an optimal spectrum reuse strategy in a heterogeneous network with in-band relay nodes to maximize the overall proportional fairness metric. We first provide one distributed resource allocation algorithm based on the alternating direction method of multipliers for the heterogeneous network with in-band relay nodes. Even though the number of feasible spectrum reuse patterns increases exponentially with the total number of base stations and relay nodes, we further prove that we can achieve the optimal network performance by only activating a limited number of spectrum reuse patterns, which is just in the order of the total number of mobile stations and relay nodes. According to this finding, we put forth an algorithm to refine the set of active reuse patterns in a soft manner by employing the re-weighted ℓ_1 -norm algorithm. Numerical simulations demonstrate the superiority and effectiveness of our proposed algorithms.

INDEX TERMS Heterogeneous network, relay node, frequency reuse, distributed algorithm, sparse optimization.

I. INTRODUCTION

Heterogeneous networks (HetNets) have been regarded as one key enabling solution in the next-generation wireless communication systems. Different from the traditional homogeneous network, a HetNet contains various low-power nodes, e.g. pico-base stations (BSs), femto-BSs, and relay nodes (RNs) [1], [2]. By deploying these low-power nodes densely, the network performance can be enhanced significantly. However, the inter-cell interference created by the nodes transmitting at different power levels has to be taken care of carefully to unleash all the potentials of HetNets. Compared with the pico-BSs and the femto-BSs, RNs are more preferred and have attracted a lot of attentions since RNs are cost-effective and can be deployed fast and timely [3].

The optimal resource allocation in a HetNet with in-band RNs is challenging since various links including the direct links (the link between a BS and a mobile station (MS)), the access links (the link between a BS and an RN), and the backhaul links (the link between an RN and an MS) all

share the same but limited spectrum resources [4]. It is very important to have an efficient resource allocation scheme to ensure satisfactory serving data rate to each MS in the HetNet with RNs.

A. RELATED WORKS

1) RESOURCE ALLOCATION IN HetNets WITHOUT RNs

A lot of works have been done to optimize the resource allocations in a HetNet without RNs [5]–[17]. Particularly, Zhang *et al.* [5] and Xue *et al.* [6] allocated orthogonal spectra among neighboring BSs to mitigate the strong inter-cell interference. In [7], spectrum reuse by allocating the same spectra to different BSs was exploited to improve the system performance. Fractional frequency reuse (FFR) has been studied by many researches. In [8], three particular FFR schemes were proposed. To exploit all feasible reuse modes, Zhuang *et al.* [9], [10] and Kuang *et al.* [11] developed the optimal resource allocation strategies. However, the centralized solutions proposed in [9]–[11] put too much computation

burden on the central controller. To offload the computation tasks from the central controller, the resource allocation and user association schemes were determined in a distributed manner in [12] and self-enforcing spectrum sharing rules were studied in [13] by formulating the problem as games. Furthermore, dynamic spectrum access (DSA) based on the principle of cognitive radio has also attracted a lot of attention [14], [15]. Comprehensive surveys on the recent advances in the resource allocations in HetNets and more references can be found in [16] and [17].

2) RESOURCE ALLOCATION IN HetNets WITH IN-BAND RNS

It is not straightforward to extend those resource allocation schemes developed for HetNets without RNs to the case with in-band RNs. This is due to the fact that the in-band RNs play two roles at the same time, i.e. role of BS and role of MS, which complicates the designs. In this paper, we focus HetNets with in-band RNs. Detailed surveys on resource allocations in HetNets with in-band RNs can be found in [18] and [19] and interested readers can refer to the references therein. Specifically, Shim *et al.* [20] proposed orthogonal subcarrier allocation among the RNs and the served MSs of one BS to avoid the strong intra-cell interference. To make full use of the limited spectrum resources and exploit the spatial reuse gains, Lee *et al.* [21] and Oyman [22] proposed to allow the RNs to reuse the same subcarriers when the resulting mutual interference is negligible. However, the schemes in [21] and [22] depended on the particular deployments of the RNs. Non-orthogonal allocations of subcarriers with an arbitrary deployment of RNs were further studied in [23]. However, the resource allocation scheme proposed in [23] assumed high-capacity wireless backhaul links and had a high implementation complexity as pointed out in [18].

Considering that in-band RNs are playing an increasingly important role in the next-generation wireless network, it is important to develop an optimal resource allocation scheme in a HetNet with in-band RNs to fully benefit from the potential performance. When studying the optimal spectrum reuse scheme in a HetNet with an arbitrary deployment of RNs, we need to consider the spectrum reuse patterns as studied in [9]–[11]. Moreover, it is preferred to establish the resource allocation scheme in a cooperative and distributed way, which exploits the computing capacities of all the nodes in the network.

B. OUR CONTRIBUTIONS

In this paper, we seek for the optimal spectrum reuse scheme for a HetNet with in-band RNs. Even though the spectrum reuse patterns for a general HetNet have been studied in [9]–[11], it is not straightforward to extend their results to a HetNet with in-band RNs. In fact, each RN in the HetNet can perform two roles simultaneously, i.e. a virtual “BS” serving others and a virtual “MS” served by others. By decoupling these two roles of each RN, we propose a novel system model and identify the optimal spectrum reuse strategy to maximize the proportional fairness (PF) metric among the data rates of

all the served MSs. Further, the resource allocation to each link is determined in a distributed fashion instead of solving the whole problem in a central controller as in [9]–[11]. Our main technical contributions in this paper are summarized as follows.

- 1) *Distributed Resource Allocation*: To offload the heavy computation burden in the central controller, we make use of the computation capacities of all the nodes, i.e. RNs and MSs, and put forth a distributed resource allocation algorithm. With the alternating direction method of multiplier (ADMM) [24], the problem is solved iteratively. In particular, we obtain closed-form solutions to update the bandwidth allocated to each spectrum reuse and association link. Furthermore, our proposed distributed algorithm can be implemented by only exchanging the local information with respect to each node, which lowers both the computation and the communication costs;
- 2) *Sparse Spectrum Reuse Scheme*: To achieve the optimal PF metric in the HetNet with RNs, each feasible spectrum reuse pattern should be exploited. However, the number of feasible spectrum reuse patterns increases exponentially with the total number of BSs and RNs. Without limiting the number of active reuse patterns, the optimal solution may activate too many reuse patterns to realize in practice. To obtain a practical solution and reduce the implementation cost, we prove that we only need to activate a small number of spectrum reuse patterns while still being able to achieve the optimal network performance. In this way, we are allowed to optimize the network with sparse spectrum reuse schemes;
- 3) *Soft Active Pattern Identification*: After enforcing the sparsity constraint on the number of active reuse patterns, we come across a non-convex problem. We employ the re-weighted ℓ_1 -norm algorithm [25] to identify the active reuse patterns in a soft manner rather than choose the active reuse patterns one by one based on the Frank-Wolfe method. This soft identification approach can accelerate the procedure of pattern selection and provide better performance than [9]–[11].

Our recent work in [26] studied the optimal time reuse in a D2D-enabled HetNet. The main technical differences between the current paper and the work in [26] are as follows.

- The distributed algorithm proposed in [26] was based on a fixed pattern set, which was decided by the central controller. In this paper, we combine the pattern selection procedure and the distributed resource allocation;
- The hard pattern identification algorithm was discussed in [26] and no sparsity constraints were enforced in the optimization. However, this paper focuses on the sparse spectrum reuse by enforcing the sparsity constraint and puts forward the soft active pattern identification scheme.

C. OUTLINE OF THE PAPER

Section II describes the system model and provides the problem formulation considering all feasible spectrum reuse patterns. In Section III, we propose one distributed resource allocation algorithm to offload the computation burden of the central controller. In Section IV, we show that a sparse spectrum reuse scheme can also achieve the optimal PF metric among the data rates of the served MSs and propose the soft active pattern identification. Section V contains numerical simulation results and Section VI concludes the paper.

D. NOTATIONS

Notations $(x)_a^b$, $\|\mathbf{x}\|_p$, \mathbf{A}^T , \mathbf{A}^{-1} , $\text{conv}(\mathcal{A})$, and $|\mathcal{A}|$ stand for $\min(b, \max(x, a))$, the ℓ_p -norm of vector \mathbf{x} , the transpose of matrix \mathbf{A} , the inverse of matrix \mathbf{A} , the convex hull of the set \mathcal{A} , and the cardinality of the set \mathcal{A} respectively. The indicator function $\mathbb{1}\{\cdot\}$ stands for the indicator function and takes the value of 1 (0) when the specified condition is met (otherwise). Notations $\mathbf{1}$ and \mathbf{I} denote the column vector with all elements being 1 and the identity matrix.

II. SYSTEM MODEL

We consider the downlink of a cooperative HetNet, where a number of MSs, RNs and conventional BSs, e.g. macro-BSs or pico-BSs, coexist as shown in Fig. 1. In this network, an MS can be served by one BS directly or fetch data from BSs through the help of one RN. The communication link between an MS and its serving BS is termed as ‘‘direct link’’. In the presence of RNs, there are two more links. One is the ‘‘access link’’, through which an MS fetches data from an RN. The other link is the ‘‘backhaul link’’, which refers to the communication link between a BS and an RN. Assume B BSs, K RNs, and M MSs exist in the HetNet and define the set of BSs, the set of RNs, and the set of MSs by $\mathcal{B} := \{1, \dots, B\}$, $\mathcal{K} := \{1, \dots, K\}$, and $\mathcal{M} := \{1, \dots, M\}$, respectively. To proceed with the system model, we make the following two assumptions:

- AS-1: Only the ‘‘half-duplexing’’ in-band RNs are considered in this paper. On the one hand, the RNs cannot transmit and receive concurrently at the same frequency [27]. On the other hand, simultaneous transmission and reception is allowed when non-overlapping spectrum resources are allocated;
- AS-2: the RNs do not receive data from other RNs, which means only one-hop RNs are considered.

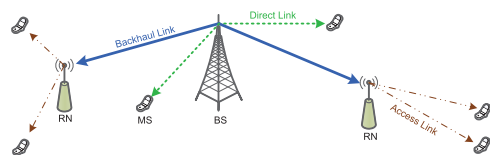


FIGURE 1. One HetNet with one BS, two RNs, and six MSs.

The RNs can be decoupled into ‘‘BSs’’ and ‘‘MSs’’ since each RN can be regarded either as one ‘‘BS’’ or as one ‘‘MS’’ depending on the function it performs. Accordingly, we define the set including both BSs and RNs as the set of servers \mathcal{N} and the set including both MSs and RNs as the set of users \mathcal{U} , i.e.

$$\mathcal{N} := \underbrace{\{1, \dots, B\}}_{\text{BSs}}, \underbrace{\{B+1, \dots, B+K\}}_{\text{RN}} \quad (1)$$

$$\mathcal{U} := \underbrace{\{1, \dots, M\}}_{\text{MSs}}, \underbrace{\{M+1, \dots, M+K\}}_{\text{RN}} \quad (2)$$

Let $N := |\mathcal{N}| = B + K$, $U := |\mathcal{U}| = M + K$. To characterize the spectrum occupation status of all servers, we employ the definition of spectrum reuse pattern proposed in [9]–[11] and denote the status of servers under spectrum reuse pattern- i by a vector $\mathbf{v}_i := [v_{i,1}, \dots, v_{i,N}]^T$, where $v_{i,n} = 1$ indicates the n -th server occupies the spectrum, and $v_{i,n} = 0$ means the n -th server does not occupy the spectrum. With the decoupling to RNs, the total number of all possible reuse patterns is $(2^N - 1)$ excluding the all 0 pattern. The set of all reuse patterns is denoted by $\mathcal{I} := \{1, \dots, 2^N - 1\}$. Furthermore, the set of active servers under pattern- i is defined as $\mathcal{A}_i := \{n | v_{i,n} = 1, \forall n \in \mathcal{N}\}$. Let x_i denote the fraction of the total available bandwidth W allocated to the spectrum reuse pattern- i . The overall spectrum reuse profile (also called reuse strategy) is thus given by the vector $\mathbf{x} := [x_1, \dots, x_{2^N-1}]^T$. Note the optimal spectrum allocation strategy should occupy the whole available bandwidth, i.e. $\sum_{i \in \mathcal{I}} x_i = 1$. Moreover, each active server under pattern- i needs to allocate the available spectrum resources to its served users. In this paper, we assume each server allocates orthogonal resources to its served users to avoid the strong intra-cell interference. Let $y_{u,n,i} \geq 0$ denote the fraction of the whole bandwidth W that is allocated to the u -th user by the n -th server under pattern- i , the overall resource allocation profile is given by the vector $\mathbf{y} := [y_{u,n,i}, \forall u, n, i]^T$. Since the available spectrum resources of each server under one pattern are bounded by the reuse pattern bandwidth and each server must make full use of its available spectrum resources, the following constraints need to be satisfied:

$$\sum_{u \in \mathcal{U}} y_{u,n,i} = \mathbb{1}\{n \in \mathcal{A}_i\} \cdot x_i, \quad \forall n \in \mathcal{N}, i \in \mathcal{I} \quad (3)$$

It should be noted that the constraints in (3) couple two variables, i.e. the reuse profile \mathbf{x} and the resource allocation profile \mathbf{y} . Given the allocated resources \mathbf{y} , the aggregated data rate of the u -th user is thus given by $R_u = W \cdot \sum_{i \in \mathcal{I}} \sum_{n \in \mathcal{N}} y_{u,n,i} \cdot c_{u,n,i}$, where $c_{u,n,i}$ denotes the spectral efficiency of the link between the n -th server and the u -th user under pattern- i . From AS-1 and AS-2, the spectral efficiencies $\{c_{u,n,i}, \forall u, n, i\}$ should be set appropriately as follows.

- Case 1: $u \leq M$. In this case, the server could be either a BS or an RN and the spectral efficiency can be

determined as

$$c_{u,n,i} = \mathbb{1}\{n \in \mathcal{A}_i\} \cdot \log_2(1 + \gamma_{u,n,i}) \text{ bits/s/Hz}, \quad (4)$$

where $\gamma_{u,n,i}$ stands for the signal-to-interference-plus-noise ratio (SINR) of the link between the u -th user and the n -th server under pattern- i . Each server in \mathcal{A}_i transmits data to the users over the assigned spectrum. The SINR $\gamma_{u,n,i}$ is thus

$$\gamma_{u,n,i} = \frac{P_n \cdot |g_{n,u}|^2}{N_0 + \sum_{m \in \mathcal{A}_i, m \neq n} P_m \cdot |g_{m,u}|^2}, \quad (5)$$

where N_0 is the power spectral density (PSD) of the thermal noise at each user, P_n denotes the PSD of the n -th server, and $g_{n,u}$ denotes the average gain of the channel from the n -th server to the u -th user. All the effects including the path-loss, the shadowing, and the fading are assumed to be absorbed into this single effective parameter;

- Case 2: $u > M$. Note the user is an RN actually, i.e. the $(u - M)$ -th RN. If the RN is active under pattern- i , it cannot receive data from the other servers according to the AS-1. Therefore, the spectral efficiency is set to be zero when the RN plays as an active server under pattern- i , i.e. $c_{u,n,i} = 0$, if $(u - M + B) \in \mathcal{A}_i$. Furthermore, only the one-hop RNs are considered as mentioned in AS-2. Thus the RN does not receive any data from the other RNs and the link between two RNs is invalid, i.e. $c_{u,n,i} = 0$, if $n > B$. The spectral efficiency $c_{u,n,i}$ in other cases can be calculated according to (4) and (5).

Note that our system model can be extended easily to full-duplex and multi-hop enabled scenarios by changing the calculation of spectral efficiencies $\{c_{u,n,i}, \forall u, n, i\}$. Denote the total throughput toward the served users from the n -th server by \tilde{R}_n . We have $\tilde{R}_n = W \cdot \sum_{i \in \mathcal{I}} \sum_{u \in \mathcal{U}} y_{u,n,i} \cdot c_{u,n,i}, \forall n \in \mathcal{N}$. Since one RN cannot generate new data to MSs, the rate of the backhaul link should be no less than the access link rate, i.e. $R_{k+M} \geq \tilde{R}_{k+B}, \forall k \in \mathcal{K}$. Furthermore, note that the overall network performance is only determined by the effective rates of the MSs. It is wasteful to allocate redundant resources to the backhaul links. The optimal resource allocation scheme must balance the resource allocated to backhaul links and access links, i.e. the optimal scheme has $R_{k+M} = \tilde{R}_{k+B}, \forall k \in \mathcal{K}$.

In summary, we can formulate the optimization problem in (6) to identify the optimal spectrum reuse strategy and the corresponding resource allocation scheme that can maximize the network PF metric. Note the fractional user association [12] is assumed in (6), namely, each user is allowed to be served by multiple servers.

Although the reuse pattern has been proposed in [9]–[11], the constraints in (6d) are not introduced, which complicates the solution to the problem. Different from solving the problem by the central controller [9]–[11], we next propose one distributed algorithm to relieve the computation burden of the central controller. Note that the noise level: N_0 and the

channel power gains from each server to the user: $\{|g(n, u)|^2\}$ need to be fed back to the central controller. Accordingly, all the spectral efficiencies of the feasible links can be derived and the problem in (6) can be solved by the central controller.

$$\max_{\mathbf{x}, \mathbf{y}} \sum_{u \in \mathcal{M}} \log(R_u) \quad (6a)$$

$$\text{s.t. } R_u = W \cdot \sum_{i \in \mathcal{I}} \sum_{n \in \mathcal{N}} y_{u,n,i} \cdot c_{u,n,i}, \quad \forall u \in \mathcal{U} \quad (6b)$$

$$\tilde{R}_n = W \cdot \sum_{i \in \mathcal{I}} \sum_{u \in \mathcal{U}} y_{u,n,i} \cdot c_{u,n,i}, \quad \forall n \in \mathcal{N} \quad (6c)$$

$$R_{k+M} = \tilde{R}_{k+B}, \quad \forall k \in \mathcal{K} \quad (6d)$$

$$\sum_{i \in \mathcal{I}} x_i = 1, \quad x_i \geq 0, \quad \forall i \in \mathcal{I} \quad (6e)$$

$$\mathbb{1}\{n \in \mathcal{A}_i\} \cdot x_i = \sum_{u \in \mathcal{U}} y_{u,n,i}, \quad \forall n \in \mathcal{N}, i \in \mathcal{I} \quad (6f)$$

$$y_{u,n,i} \geq 0, \quad \forall u \in \mathcal{U}, n \in \mathcal{N}, i \in \mathcal{I} \quad (6g)$$

III. DISTRIBUTED RESOURCE ALLOCATION

Since the objective function (6a) is concave and all the constraints are affine, we have a convex optimization problem in (6) and the optimal solution can be obtained by using a standard package such as CVX [28]. However, the number of all feasible spectrum reuse patterns increases exponentially with the number of servers. It becomes very hard to solve the problem at the central controller even for a HetNet of a medium size [9], [11]. In order to relieve the computation burden of the central controller and fully exploit the computation capacity of the cooperative HetNet, we develop a distributed resource allocation algorithm to distribute computation tasks to all the nodes including RNs and MSs.

As shown in (6f), the balance between the reuse pattern bandwidth x_i and the total resource allocated to the served users of the n -th server under pattern- i , i.e. $\sum_{u \in \mathcal{U}} y_{u,n,i}$, is important. To address this balance, we employ the method of multipliers and solve the problem in (6) iteratively. Let \mathcal{X} denote the feasible region of \mathbf{x} determined by (6e) and \mathcal{Y} denote the feasible region of \mathbf{y} determined by (6d) and (6g), i.e. $\mathcal{X} := \{\mathbf{x} | \sum_{i \in \mathcal{I}} x_i = 1; x_i \geq 0, \forall i \in \mathcal{I}\}$, $\mathcal{Y} := \{\mathbf{y} | R_{k+M} = \tilde{R}_{k+B}, \forall k \in \mathcal{K}; y_{u,n,i} \geq 0, \forall u \in \mathcal{U}, n \in \mathcal{N}, i \in \mathcal{I}\}$. Then the augmented Lagrangian for the problem in (6) can be formulated as

$$\begin{aligned} \mathcal{L}_{\rho_1}(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) = & \sum_{u \in \mathcal{M}} \log(W \sum_{n \in \mathcal{N}} \sum_{i \in \mathcal{I}} y_{u,n,i} c_{u,n,i}) \\ & + \sum_{n \in \mathcal{N}} \sum_{i \in \mathcal{I}} \lambda_{n,i} (\mathbb{1}\{n \in \mathcal{A}_i\} x_i - \sum_{u \in \mathcal{U}} y_{u,n,i}) \\ & - \frac{\rho_1}{2} \sum_{n \in \mathcal{N}} \sum_{i \in \mathcal{I}} (\mathbb{1}\{n \in \mathcal{A}_i\} x_i - \sum_{u \in \mathcal{U}} y_{u,n,i})^2, \end{aligned} \quad (7)$$

where $\mathbf{x} \in \mathcal{X}$, $\mathbf{y} \in \mathcal{Y}$, $\boldsymbol{\lambda} := [\lambda_{n,i}, \forall n, i]^T$ is the Lagrange multiplier (a.k.a dual variable), and ρ_1 is the penalty coefficient. Note the feasible region \mathcal{Y} has nothing to do with

the constraint in (6f), which is taken into account in the above augmented Lagrangian. During the t -th iteration of the method of multipliers, the variables, i.e. \mathbf{x} , \mathbf{y} , and $\boldsymbol{\lambda}$, can be updated as

$$\mathbf{x}^{(t+1)} = \arg \max_{\mathbf{x} \in \mathcal{X}} \mathcal{L}_{\rho_1}(\mathbf{x}, \mathbf{y}^{(t)}, \boldsymbol{\lambda}^{(t)}); \quad (8a)$$

$$\mathbf{y}^{(t+1)} = \arg \max_{\mathbf{y} \in \mathcal{Y}} \mathcal{L}_{\rho_1}(\mathbf{x}^{(t+1)}, \mathbf{y}, \boldsymbol{\lambda}^{(t)}); \quad (8b)$$

$$\lambda_{n,i}^{(t+1)} = \lambda_{n,i}^{(t)} - \rho_1 (\mathbb{1}\{n \in \mathcal{A}_i\} x_i^{(t+1)} - \sum_{u \in \mathcal{U}} y_{u,n,i}^{(t+1)}). \quad (8c)$$

Note the primal updates in (8a) and (8b) are solved to update the reuse pattern bandwidth $\{x_i, \forall i\}$ and link bandwidth $\{y_{u,n,i}, \forall u, n, i\}$, respectively. The primal updates, i.e. the *Reuse Pattern Bandwidth Update* in (8a) and the *Link Bandwidth Update* in (8b) can be solved easily in closed-form. Next, we give the closed-form of the primal-dual updates in (8). Detailed derivations can be found in Appendix 1.

A. REUSE PATTERN BANDWIDTH UPDATE

In the t -th iteration, the bandwidths of reuse patterns are updated as

$$x_i^{(t+1)} = \left[\frac{\sum_n \mathbb{1}\{n \in \mathcal{A}_i\} (\lambda_{n,i}^{(t)} + \rho_1 \sum_u y_{u,n,i}^{(t)}) + \theta}{\rho_1 \sum_n \mathbb{1}\{n \in \mathcal{A}_i\}} \right]_0, \quad (9)$$

where θ is chosen such that $\sum_{i \in \mathcal{I}} x_i^{(t+1)} = 1$. The central controller is employed to execute this task since the update needs information including the link bandwidths $\{y_{u,n,i}^{(t)}, \forall u, n, i\}$ and dual variables $\{\lambda_{n,i}^{(t)}, \forall n, i\}$. Note the update has the closed-form solution and the computation complexity is low.

B. LINK BANDWIDTH UPDATE

When updating the link bandwidths, due to the large number of variables, i.e. $\{y_{u,n,i}, \forall u, n, i\}$, it is necessary to develop a distributed algorithm. One popular solution is to employ the ADMM algorithm [24]. To this end, we can introduce an auxiliary variable \mathbf{z} and let $\mathbf{z} = \mathbf{y}$. Then the problem in (8b) becomes

$$\begin{aligned} \max_{\mathbf{y} \geq \mathbf{0}, \mathbf{z}} \quad & \sum_{u \in \mathcal{M}} \log(W \sum_{i \in \mathcal{I}} \sum_{n \in \mathcal{N}} y_{u,n,i} \cdot c_{u,n,i}) \\ & + \sum_{n \in \mathcal{N}} \sum_{i \in \mathcal{I}} \lambda_{n,i}^{(t)} (\mathbb{1}\{n \in \mathcal{A}_i\} x_i^{(t+1)} - \sum_{u \in \mathcal{U}} z_{u,n,i}) \\ & - \frac{\rho_1}{2} \sum_{n \in \mathcal{N}} \sum_{i \in \mathcal{I}} (\mathbb{1}\{n \in \mathcal{A}_i\} x_i^{(t+1)} - \sum_{u \in \mathcal{U}} z_{u,n,i})^2 \end{aligned} \quad (10a)$$

$$\text{s. t. } R_u = W \sum_{i \in \mathcal{I}} \sum_{n \in \mathcal{N}} y_{u,n,i} \cdot c_{u,n,i}, \quad \forall u \in \mathcal{U}, \quad (10b)$$

$$\tilde{R}_n = W \sum_{i \in \mathcal{I}} \sum_{u \in \mathcal{U}} z_{u,n,i} \cdot c_{u,n,i}, \quad \forall n \in \mathcal{N} \quad (10c)$$

$$R_{k+M} = \tilde{R}_{k+B}, \quad \forall k \in \mathcal{K}, \quad (10d)$$

$$y_{u,n,i} = z_{u,n,i}, \quad \forall u \in \mathcal{U}, n \in \mathcal{N}, i \in \mathcal{I}. \quad (10e)$$

Note that the variable \mathbf{y} can be regarded as the link bandwidth requests from users and the variable \mathbf{z} can be regarded as the link bandwidths decisions made by the servers in the above problem. Based on the ADMM principle, we can develop a distributed algorithm to solve the optimization problem in (10). In the rest of this subsection, we just present the update rules for the primal variables: \mathbf{y} , \mathbf{z} and the dual variables: $\boldsymbol{\alpha}$, $\boldsymbol{\xi}$ in the j -th iteration of the ADMM. Details about the augmented Lagrangian, the definition of the dual variables $\boldsymbol{\alpha}$, $\boldsymbol{\xi}$, and the corresponding derivations are outlined in Appendix 1-B.

1) LINK BANDWIDTH REQUEST UPDATE

At the u -th user, the link bandwidth requests transmitted to all servers under all patterns, i.e. variables $\{y_{u,n,i}, \forall n, i\}$, are updated as

$$y_{u,n,i}^{(j+1)} = \left[z_{u,n,i}^{(j)} + \frac{W \cdot c_{u,n,i} \cdot \Gamma_{u,n,i}^{(j)} - \xi_{u,n,i}^{(j)}}{\rho_3} \right]_0, \quad (11)$$

where $\xi_{u,n,i}^{(j)}$ and ρ_3 are the Lagrange multiplier and penalty coefficient associated with the constraint $z_{u,n,i} = y_{u,n,i}$ in the j -th iteration. Note $\Gamma_{u,n,i}^{(j)}$ is defined as

$$\begin{aligned} \Gamma_{u,n,i}^{(j)} := & \frac{\mathbb{1}\{u \leq M\}}{R_u} + \mathbb{1}\{u > M\} \\ & \cdot \left(\alpha_{u-M}^{(j)} - \rho_2 \cdot \left(R_u - \tilde{R}_{u-M+B} \right) \right), \end{aligned} \quad (12)$$

where $\{\alpha_k^{(j)}, \forall k\}$ and ρ_2 are the Lagrange multipliers and penalty coefficient associated with the constraints in (6d) in the j -th iteration.

2) LINK BANDWIDTH DECISION UPDATE

At the n -th server, the link bandwidth decisions sent to all users under all patterns, i.e. variables $\{z_{u,n,i}, \forall u, i\}$, are updated as

$$z_{n,i}^{(j+1)} = \mathbf{P} \cdot \left(\mathbf{b}_{n,i}^{(j)} - \mathbb{1}\{n > B\} \cdot W \cdot \boldsymbol{\Lambda}_{n,i}^{(j)} \right), \quad (13)$$

where the matrix $\mathbf{P} := (\rho_1 \mathbf{1}\mathbf{1}^T + \rho_3 \mathbf{I})^{-1}$, the vector $\mathbf{z}_{n,i}^{(j+1)} := [z_{1,n,i}^{(j+1)}, \dots, z_{U,n,i}^{(j+1)}]^T$, the vector $\mathbf{b}_{n,i}^{(j)} := [b_{1,n,i}^{(j)}, \dots, b_{U,n,i}^{(j)}]^T$, where each element of $\mathbf{b}_{n,i}^{(j)}$ is defined in (36), and the vector $\boldsymbol{\Lambda}_{n,i}^{(j)}$ is defined as

$$\boldsymbol{\Lambda}_{n,i}^{(j)} := \left(\alpha_{n-B}^{(j)} - \rho_2 \cdot \left(R_{n-B+M} - \tilde{R}_n \right) \right) \cdot \mathbf{c}_{n,i}, \quad (14)$$

where $\mathbf{c}_{n,i} := [c_{1,n,i}, \dots, c_{U,n,i}]^T$.

3) DUAL VARIABLE UPDATE

At the n -th server, the dual variables are updated as

$$\alpha_k^{(j+1)} = \alpha_k^{(j)} - \rho_2 \cdot \left(R_{k+M} - \tilde{R}_{k+B} \right), \quad \forall k; \quad (15)$$

$$\xi_{u,n,i}^{(j+1)} = \xi_{u,n,i}^{(j)} - \rho_3 \cdot \left(z_{u,n,i}^{(j+1)} - y_{u,n,i}^{(j+1)} \right), \quad \forall u, i. \quad (16)$$

The above iterations including the ‘‘link bandwidth request update’’ in (11), the ‘‘link bandwidth decision update’’

TABLE 1. Information exchanges in Algorithm 1.

Node	Available Information	Received Information	Transmitted Information
u -th User	$c_{u,n,i}, \forall n, i; y_{u,n,i}, \forall n, i$	$\alpha_k, \forall k; \xi_{u,n,i}, \forall n, i; z_{u,n,i}, \forall n, i$	$y_{u,n,i}, \forall n, i$
n -th Server	$c_{u,n,i}, \forall u, i; \mathbb{1}\{n > B\}\alpha_{n-B};$ $\xi_{u,n,i}, \forall u, i; z_{u,n,i}, \forall u, i; \lambda_{n,i}, \forall i$	$y_{u,n,i}, \forall u, i; x_i, \forall i$	$\mathbb{1}\{n > B\}\alpha_{n-B}; \xi_{u,n,i}, \forall u, i;$ $z_{u,n,i}, \forall u, i; \lambda_{n,i}, \forall i$
Central Controller	Null	$y_{u,n,i}, \forall u, n, i; \lambda_{n,i}, \forall n, i$	$x_i, \forall i$

in (13), and the “dual variable updates” in (15) and (16) should be carried out in a distributed fashion until convergence. With one spectrum reuse profile $\mathbf{x}^{(t+1)}$ as the input to the update in (8b), the solution to the problem in (10), i.e. the converging variables \mathbf{y} and \mathbf{z} , serves as the link bandwidth update $\mathbf{y}^{(t+1)}$ in (8b). The constraints in (10) will ensure $\mathbf{y}^{(t+1)}$ is feasible, i.e. $\mathbf{y}^{(t+1)} \in \mathcal{Y}$.

C. DUAL UPDATE

During the t -th iteration, the Lagrange multipliers $\{\lambda_{n,i}, \forall i\}$ are updated at the n -th server as

$$\lambda_{n,i}^{(t+1)} = \lambda_{n,i}^{(t)} - \rho_1 (\mathbb{1}\{n \in \mathcal{A}_i\} x_i^{(t+1)} - \sum_{u \in \mathcal{U}} y_{u,n,i}^{(t+1)}). \quad (17)$$

The Lagrange multipliers can be regarded as the prices in the network. Specifically, when the available resources of the n -th server under pattern- i are abundant, i.e. $x_i^{(t+1)} > \sum_{u \in \mathcal{U}} y_{u,n,i}^{(t+1)}$, the n -th server reduces its price to encourage its users to utilize more resources under pattern- i .

D. ALGORITHM SUMMARY

Compared to solving the problem in (6) by some standard solvers, e.g. CVX [28], the distributed solution offloads the computation tasks to neighboring nodes. Algorithm 1 summarizes the steps of our proposed distributed resource allocation and the corresponding information exchanges in the HetNet are shown in Table 1. Although a distributed algorithm has been developed, global information of the HetNet is needed at each node to update the variables. However, each server only has a limited coverage due to the finite transmit power. This indicates that some variables in the link bandwidths do not need to be exchanged between the servers and the users. In fact, only the local information is needed for *Link Bandwidth Update* at each node. Specifically, we let $\mathcal{N}_{u,i}$ denote the set of servers under pattern- i whose spectral efficiency to the u -th user is above the threshold c_0 , i.e. $\mathcal{N}_{u,i} := \{n | c_{u,n,i} > c_0, \forall n \in \mathcal{N}, i \in \mathcal{I}\}$. For the u_0 -th user, it only needs to update the link bandwidth requests $\{y_{u_0,n,i}, \forall n \in \mathcal{N}_{u_0,i}, i \in \mathcal{I}\}$ and send to the corresponding servers. Accordingly, for the n_0 -th server, it only needs to make the link bandwidth decisions $\{z_{u,n_0,i}, \forall u \in \mathcal{U}_{n_0,i}, i \in \mathcal{I}\}$, where $\mathcal{U}_{n_0,i} := \{u | \exists n_0 \in \mathcal{N}_{u,i}, \forall u \in \mathcal{U}, i \in \mathcal{I}\}$. Then the updates in Algorithm 1 can be modified to updating with only local information exchanges, which results in lower computation and communication costs. This cost reduction is especially evident when the number of RNs is large. In Section V, we will verify the performance degradation due to local information exchanges is negligible with numerical results.

Algorithm 1 Central Update of Reuse Bandwidths & Distributed Update of Link Bandwidths

- 1: **INPUT:** The active reuse pattern set: \mathcal{I} and the corresponding feasible set of the reuse profiles: $\mathcal{X} := \{\mathbf{x} | \sum_{i \in \mathcal{I}} x_i = 1; x_i \geq 0, \forall i \in \mathcal{I}\}$;
- 2: **Initialization:** Initialize variables: $\rho_1, \rho_2, \rho_3, \mathbf{x}^{(1)}, \mathbf{y}^{(1)}, \boldsymbol{\lambda}^{(1)}, \boldsymbol{\alpha}^{(1)}, \boldsymbol{\xi}^{(1)}$;
- 3: Initialize $t = 1$;
- 4: **Repeat**
- 5: *Central Controller-Side:*
- 6: Update the reuse pattern bandwidths \mathbf{x} as (9);
- 7: Initialize $j = 1$;
- 8: **Repeat**
- 9: *User-Side:*
- 10: Update link bandwidth request \mathbf{y} as (11);
- 11: *Server-Side:*
- 12: Update link bandwidth decision \mathbf{z} as (13);
- 13: Update $\boldsymbol{\alpha}$ and $\boldsymbol{\xi}$ as (15) and (16);
- 14: $j = j + 1$;
- 15: **Until** termination criterion is satisfied;
- 16: *Server-Side:*
- 17: Update Lagrange multiplier $\boldsymbol{\lambda}$ as (17);
- 18: $t = t + 1$;
- 19: **Until** termination criterion is satisfied;
- 20: **OUTPUT:** the optimal spectrum resource allocation scheme $(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})$.

IV. SPARSE SPECTRUM REUSE SCHEME

In Section III, we have developed one distributed resource allocation solution to the problem in (6). However, the reuse profile update in (9) could end up with a reuse profile with all non-zero entries. In other words, all possible reuse patterns are activated. Recall that the total number of reuse patterns increases exponentially with N . It is impractical to implement the reuse profile. To circumvent this dilemma, we next discuss how to activate as few reuse patterns as possible while still being able to achieve a satisfactory network performance.

A. UPPER BOUND ON THE NUMBER OF ACTIVE SPECTRUM REUSE PATTERNS

Before we put forth algorithms to pursue sparse spectrum reuse profiles, we first establish the following result to show that the optimal network PF metric indeed can be achieved with a sparse spectrum reuse scheme. The proof is given in Appendix 2.

Proposition 1: There exists one optimal solution \mathbf{x}^ to the problem in (6) where at most $(M + K)$ out of $(2^N - 1)$ reuse*

patterns are active. Specifically, we can have one optimal reuse profile satisfying the condition: $\|\mathbf{x}^*\|_0 \leq M + K$.

When the total number of servers is large, we have $M + K \ll 2^N - 1$. Therefore, Proposition 1 indicates that the optimal performance can be realized by a sparse spectrum reuse. Intuitively, when a large number of servers coexist in the HetNet, neighboring servers observe strong interference under some particular reuse patterns. Allocating spectrum resources to these reuse patterns will degrade the network performance.

B. SOFT ACTIVE PATTERN IDENTIFICATION (SAPI)

In the above subsection, we have proved that the optimal PF metric can be achieved with a sparse reuse profile. Relying on this fact, the nonlinear column generation [29] can be invoked to select suitable reuse patterns one by one. The column generation method is also interpreted as the Frank-Wolfe method in [30]. Comparing with this widely used method, we are more interested in seeking a structure as sparse as possible. Therefore, different from the schemes in [9]–[11] and [26] where the candidate pattern set was expanded iteratively, we incorporate the sparsity constraint into the reuse pattern bandwidth update in (8a). Then the optimization problem in (6) becomes the following non-convex one:

$$\max_{\mathbf{x}, \mathbf{y}} f(\mathbf{y}) = \sum_{u \in \mathcal{M}} \log(R_u) \quad (18a)$$

$$\text{s.t. (6b) – (6g), } \|\mathbf{x}\|_0 \leq D, \quad (18b)$$

where D represents an upper bound on the number of active spectrum reuse patterns. Due to Proposition 1, it can be seen that the solution to the problem in (18) indeed achieves the optimal network performance in (6) when $D \geq M + K$.

The Lagrangian for the problem in (18) can be expressed as in (19), as shown at the bottom of this page, where $\mathbf{x} \in \mathcal{X}$, $\mathbf{y} \in \mathcal{Y}$, and $\kappa \geq 0$ is the Lagrange multiplier associated with the sparsity constraint in (18b). During the t -th iteration, similar to the updates in (8), we have the following iterative ADMM updates:

$$\mathbf{x}^{(t+1)} = \arg \max_{\mathbf{x} \in \mathcal{X}} \tilde{\mathcal{L}}_{\rho_1}(\mathbf{x}, \mathbf{y}^{(t)}, \boldsymbol{\lambda}^{(t)}, \kappa^{(t)}); \quad (20a)$$

$$\mathbf{y}^{(t+1)} = \arg \max_{\mathbf{y} \in \mathcal{Y}} \tilde{\mathcal{L}}_{\rho_1}(\mathbf{x}^{(t+1)}, \mathbf{y}, \boldsymbol{\lambda}^{(t)}, \kappa^{(t)}); \quad (20b)$$

$$\kappa^{(t+1)} = \kappa^{(t)} - \delta_\kappa \cdot (D - \|\mathbf{x}\|_0); \quad (20c)$$

$$\lambda_{n,i}^{(t+1)} = \lambda_{n,i}^{(t)} - \rho_1 (\mathbb{1}\{n \in \mathcal{A}_i\} x_i^{(t+1)} - \sum_{u \in \mathcal{U}} y_{u,n,i}^{(t+1)}), \quad (20d)$$

$$\tilde{\mathcal{L}}_{\rho_1}(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}, \kappa) = \sum_{u \in \mathcal{M}} \log(W \sum_{n \in \mathcal{N}} \sum_{i \in \mathcal{I}} y_{u,n,i} c_{u,n,i}) + \sum_{n \in \mathcal{N}} \sum_{i \in \mathcal{I}} \lambda_{n,i} (\mathbb{1}\{n \in \mathcal{A}_i\} x_i - \sum_{u \in \mathcal{U}} y_{u,n,i}) - \frac{\rho_1}{2} \sum_{n \in \mathcal{N}} \sum_{i \in \mathcal{I}} (\mathbb{1}\{n \in \mathcal{A}_i\} \cdot x_i - \sum_{u \in \mathcal{U}} y_{u,n,i})^2 + \kappa \cdot (D - \|\mathbf{x}\|_0) \quad (19)$$

$$\tilde{\mathcal{L}}_{\rho_1}(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}, \kappa) = \sum_{u \in \mathcal{M}} \log(W \sum_{n \in \mathcal{N}} \sum_{i \in \mathcal{I}} y_{u,n,i} c_{u,n,i}) + \sum_{n \in \mathcal{N}} \sum_{i \in \mathcal{I}} \lambda_{n,i} (\mathbb{1}\{n \in \mathcal{A}_i\} x_i - \sum_{u \in \mathcal{U}} y_{u,n,i}) - \frac{\rho_1}{2} \sum_{n \in \mathcal{N}} \sum_{i \in \mathcal{I}} (\mathbb{1}\{n \in \mathcal{A}_i\} \cdot x_i - \sum_{u \in \mathcal{U}} y_{u,n,i})^2 + \kappa \cdot (D - \sum_{i \in \mathcal{I}} w_i^{(t)} \cdot x_i) \quad (21)$$

where δ_κ denotes the step size in updating κ .

Due to the ℓ_0 -norm in (19), the maximization with respect to \mathbf{x} in (20a) is non-convex. A common alternative is to utilize the ℓ_1 -norm as a proxy for the ℓ_0 sparsity count and solve the resulting convex problem with the LASSO [31]. However, LASSO does not work here due to the fact that $\|\mathbf{x}\|_1 = 1$ according to the constraint in (6e). To circumvent this dilemma, we employ the re-weighted ℓ_1 -norm algorithm [25]. This algorithm gives each element a positive weight first and then relaxes the ℓ_0 -norm to the weighted ℓ_1 -norm to make the problem convex again. The re-weighted ℓ_1 -norm algorithm relax the ℓ_0 -norm in (20a) and updates the weights iteratively till convergence. Specifically, in the t -th iteration, the ℓ_0 -norm of the reuse profile is approximated by $\sum_{i \in \mathcal{I}} w_i^{(t)} \cdot x_i$ and the Lagrangian in (19) becomes the one in (21), as shown at the bottom of this page, where $w_i^{(t)}$ is the weight associated with the pattern- i in the t -th iteration. It is determined by

$$w_i^{(t)} = (x_i^{(t)} + \epsilon)^{-1}, \quad \forall i \in \mathcal{I}, \quad (22)$$

where ϵ is a small constant to prevent the weights from going to infinity. The introduction of sparsity constraint in (18b) does not affect the updates in (20b) and (20d). The modified primal-dual updates in (20a) and (20c) are given as follows. Appendix 3 has the derivation outlines.

1) PRIMAL UPDATE IN (20a)

The reuse pattern bandwidths in the t -th iteration are updated as follows.

$$x_i^{(t+1)} = \left[\frac{\sum_n \mathbb{1}\{n \in \mathcal{A}_i\} (\lambda_{n,i}^{(t)} + \rho_1 \sum_u y_{u,n,i}^{(t)}) - \kappa^{(t)} w_i^{(t)} + \theta}{\rho_1 \sum_n \mathbb{1}\{n \in \mathcal{A}_i\}} \right]_0, \quad (23)$$

where θ is chosen to ensure $\sum_{i \in \mathcal{I}} x_i^{(t+1)} = 1$.

2) DUAL UPDATE IN (20c)

The Lagrange multiplier κ in the t -th iteration is updated as

$$\kappa^{(t+1)} = \left[\kappa^{(t)} - \delta_\kappa \cdot (D - \sum_{i \in \mathcal{I}} w_i^{(t+1)} \cdot x_i^{(t+1)}) \right]_0. \quad (24)$$

Note it is important to choose an appropriate δ_κ in updating κ as explained in Appendix 3.

Compared to the hard pattern selection rules in [9]–[11] and [26], our sparse reuse scheme tends to select patterns in a soft manner. The Soft Active Pattern Identification (SAPI) is summarized in Algorithm 2. In particular, we substitute the update rule of \mathbf{x} in (23) for the update in Step 6 of Algorithm 1. Next, given the updated reuse profile $\mathbf{x}^{(t+1)}$, the sparse spectrum reuse scheme in Algorithm 2 borrows Steps 7–17 from Algorithm 1 to update the remaining variables \mathbf{y} and \mathbf{z} . The main complexity of both Algorithm 1 and Algorithm 2 lies in updating variables \mathbf{y} and \mathbf{z} as in (11) and (13). Specifically, in the j -th iteration, updating \mathbf{y} leads to the complexity of $\mathcal{O}(N2^N \log(S))$ and updating \mathbf{z} leads to the complexity of $\mathcal{O}(U2^N \log(S))$, where $\log(S)$ denotes the complexity of the bisection method which is utilized to find $R_u^{(j+1)}$ and $\tilde{R}_n^{(j+1)}$. Unlike the hard pattern selection methods, SAPI is able to find a sparse spectrum reuse scheme with controllable sparsity.

Algorithm 2 Soft Active Pattern Identification (SAPI)

- 1: **Initialization:** Initialize variables: $\rho_1, \rho_2, \rho_3, \mathbf{x}^{(1)}, \mathbf{y}^{(1)}, \boldsymbol{\lambda}^{(1)}, \boldsymbol{\alpha}^{(1)}, \boldsymbol{\xi}^{(1)}, \kappa^{(1)}, \delta_\kappa$, and ϵ ;
 - 2: Initialize $t = 1$;
 - 3: **Repeat**
 - 4: *Central Controller-Side:*
 - 5: Update the reuse pattern bandwidths \mathbf{x} as (23);
 - 6: Update the weights \mathbf{w} as (22);
 - 7: Update the Lagrange multiplier κ as (24);
 - 8: Run Step 7–17 in Algorithm 1 with input \mathcal{I} and the updated reuse profile $\mathbf{x}^{(t+1)}$;
 - 9: $t = t + 1$;
 - 10: **Until** termination criterion is satisfied;
 - 11: **OUTPUT:** the sparse reuse profile and the corresponding link bandwidths $(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})$.
-

V. NUMERICAL RESULTS

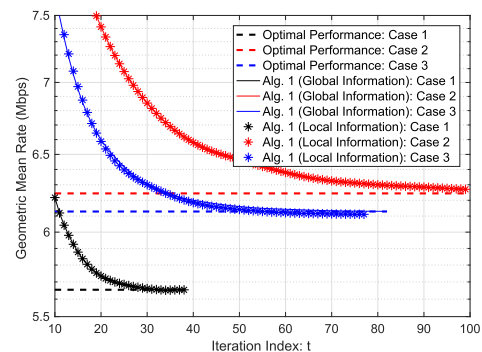
In this section, our proposed algorithms are tested by simulating HetNets with the following configured system parameters:

- Transmission power of each BS (RN) is 46dBm (30dBm) [32];
- One BS is placed at the center of a square specified by $[0, 1000]\text{m} \times [0, 1000]\text{m}$. A number of RNs and MSs are uniformly dropped in this square. All nodes are equipped with a single antenna for transmission and reception;
- System bandwidth is 20MHz and the noise PSD is -174dBm/Hz ;
- Distance-dependent path-losses of the channels from BSs to RNs ($-20 \log_{10} |g_{n,u}|, n \leq B, u > M$) and from BS/RNs to MSs ($-20 \log_{10} |g_{n,u}|, u \leq M$) are modeled as: $23.5 \log_{10} d_m + 34.5(\text{dB})$ and $35.7 \log_{10} d_m + 33.4(\text{dB})$ respectively, where d_m represents the distance in meters [32];
- Three simulated HetNet scenarios:
 - Case 1: 1 BS, 3 RNs, and 30 MSs;

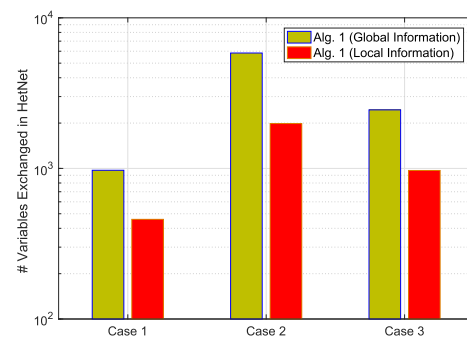
- Case 2: 1 BS, 5 RNs, and 30 MSs;
- Case 3: 2 BSs, 3 RNs, and 30 MSs;
- The network performance is measured by the geometric mean (GM) rate of the served MSs. Specifically, the GM rate is defined as $(\prod_{u \in \mathcal{M}} R_u)^{1/M}$.

A. CONVERGENCE OF ALGORITHM 1

The convergence behavior of Algorithm 1 is plotted in Fig. 2 for three different simulation scenarios. Due to the fact that the equality constraints in (6f) are not satisfied at the beginning, the initial GM rates are actually not feasible. We also plot the convergence behavior of Algorithm 1 with local information exchanges as well. The standard solver CVX is employed to obtain the optimal performance considering all possible reuse patterns. From Fig. 2, we see Algorithm 1 with either global or local information exchanges converges fast within dozens of iterations. As mentioned in Section III-D, the amount of variables that need to be exchanged among neighboring nodes can be reduced with local information exchanges as illustrated in Table 2. In Fig. 2(b), we compare the number of variables that need to be exchanged between the global information exchange scheme and the local information exchange one. It is clear that Algorithm 1



(a)



(b)

FIGURE 2. Optimal Performance: Solution to the problem in (6) with the CVX solver; Alg. 1 (Global Information): Solution to the problem in (6) obtained from Algorithm 1 with global information exchange scheme; Alg. 1 (Local Information): Solution to the problem in (6) obtained from Algorithm 1 with local information exchange scheme (c_0 is set to 0.5bits/s/Hz as described in Section III-D). (a) Convergence behavior of Algorithm 1. (b) Global vs. Local information exchanges.

TABLE 2. Comparisons between two information exchange schemes.

Node	Global Information Exchanges	Local Information Exchanges
u -th User	$y_{u,n,i}, \forall n, i$	$y_{u,n,i}, \forall n \in \mathcal{N}_{u,i}, i$
n -th Server	$\mathbb{1}\{n > B\} \alpha_{n-B}; \xi_{u,n,i}, \forall u, i; z_{u,n,i}, \forall u, i;$ $\lambda_{n,i}, \forall i$	$\mathbb{1}\{n > B\} \alpha_{n-B}; \xi_{u,n,i}, \forall u \in \mathcal{U}_{n,i}, i;$ $z_{u,n,i}, \forall u \in \mathcal{U}_{n,i}, i; \lambda_{n,i}, \forall i$
Central Controller	$x_i, \forall i$	$x_i, \forall i$

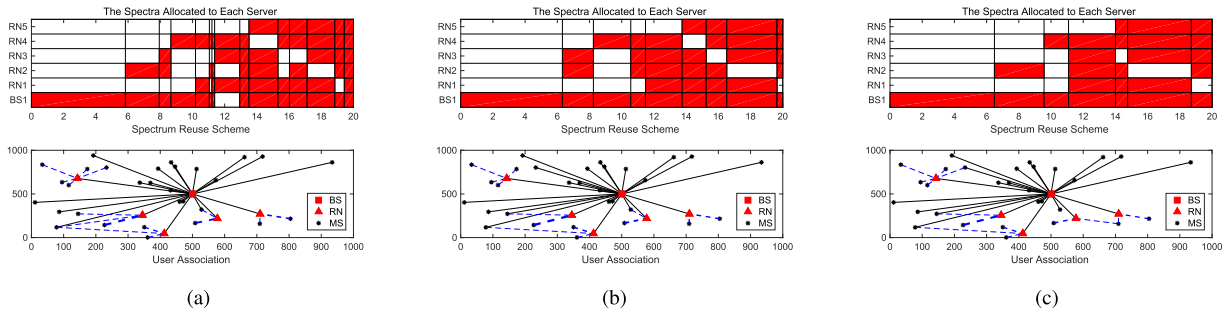


FIGURE 3. Solid color indicates that the particular spectrum is occupied by the corresponding server. The line connecting two nodes indicates a served communication link. Solid line (Dashed Line) represents the direct or backhaul link (Access Link). All the results are simulated with the HetNet in Case 2. (a) Solution-1: GM = 6.2706Mbps. (b) Solution-2: GM = 6.2294Mbps. (c) Solution-3: GM = 6.1401Mbps.

with local information exchanges reduces the communication overheads, which is especially evident for a large network.

B. SPARSE SPECTRUM REUSE STRATEGY

To show the effects of the sparsity condition in (18b), we test Algorithm 1 and Algorithm 2 by simulating the HetNet in Case 2. In Fig. 3(a), without enforcing the sparsity constraint, we see that the number of active reuse patterns in the optimal reuse strategy (Solution-1) is 15. In Fig. 3(b) and Fig. 3(c), after introducing the sparsity constraint in (18b), we simulate the HetNet with $D = 10$ and $D = 7$, respectively. From these figures, we see the performance degradation is negligible even though only a small number of reuse patterns are activated. Fig. 3 also indicates that the number of active reuse patterns, i.e. the sparsity in the reuse profile, can be well controlled with our proposed SAPI.

Fig. 4 shows that the numbers of active reuse patterns in Solution-2 and Solution-3 through SAPI are much smaller than that in Solution-1. This verifies that Algorithm 2 can obtain a sparser reuse profile than Algorithm 1 and the

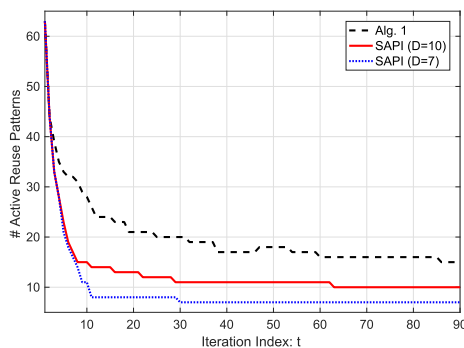


FIGURE 4. Alg. 1: Solution obtained from Algorithm 1; SAPI ($D = 10$): Solution obtained from Algorithm 2 with $D = 10$; SAPI ($D = 7$): Solution obtained from Algorithm 2 with $D = 7$. All the results are simulated with the HetNet in Case 2.

pattern selection is accelerated with the re-weighted ℓ_1 -norm algorithm. This is due to that, as the number of active spectrum reuse patterns does not satisfy the sparsity constraint, the Lagrange multiplier κ in (24) grows up. A large nonnegative κ will amplify the difference between the weights of any two reuse patterns as in (23). Due to the feasible region \mathcal{X} , the amplified difference among all feasible reuse patterns will drive some small elements in x to zero and the number of active reuse patterns gets further reduced. This also testifies the re-weighting technique can accelerate the selection of the active reuse patterns [33].

C. COMPARISONS OF ACTIVE PATTERN IDENTIFICATION SCHEMES

In Fig. 5, we simulate the HetNet in Case 3. In particular, we solve the problem in (18) with two different active pattern identification schemes, i.e. the hard pattern selection method proposed by Kuang et al. [11] and the SAPI with different values of D . Note the hard pattern selection method has

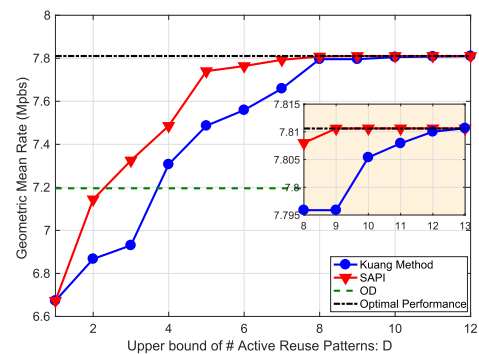


FIGURE 5. Kuang Method: Solution with the hard active pattern identification scheme proposed in [11]; SAPI: Solution with Algorithm 2; OD: Solution with the reuse strategy proposed in [34]; Optimal Performance: Solution to the problem in (6) with the CVX solver. All the results are simulated with the HetNet in Case 3.

been briefed at the beginning of Section IV-B. A similar method was also adopted in our prior work in [26], which can be regarded as the state of the art. In Fig. 5, we plot the corresponding GM rates of the MSs and we can see the achieved GM rates with either active pattern identification scheme converge to the optimal performance when $D = 13$. However, the SAPI performs better than the method proposed in [11] when D is less than 13. Note that Proposition 1 tells that the number of needed active reuse patterns is bounded by $M + K = 33$. From Fig. 5 we see 9 active patterns are sufficient to approach the optimal network performance. Furthermore, we also plot the GM rate obtained by the method in [34], where the set \mathcal{I} only contains four reuse patterns, i.e. $[1, 0, 0, 0, 0]^T$, $[0, 1, 0, 0, 0]^T$, $[1, 1, 0, 0, 0]^T$, and $[0, 0, 1, 1, 1]^T$, and serves as the input in Algorithm 1. Clearly, the simple reuse strategy in [34] cannot benefit from the full potentials of RNs.

VI. CONCLUSIONS

In this paper, we have studied the optimal spectrum reuse strategy in a HetNet with in-band RNs. We have proposed a distributed resource allocation algorithm to fully exploit the computing capacities of different nodes in the cooperative HetNet. To reduce the communication and computation cost further, we have also refined the resource allocation algorithm by only exchanging the local information among neighboring nodes. Although the number of all feasible reuse patterns increases exponentially with the total number of BSs and RNs, we have proved that the optimal network performance can be achieved by a sparse spectrum reuse strategy. Basing on this observation, we have also put forth one active pattern identification schemes, i.e the SAPI. In particular, the SAPI is based on the re-weighted ℓ_1 -norm algorithm and is able to identify the active reuse patterns in a soft manner with fewer iterations. Numerical results corroborate our designs and demonstrate our proposed spectrum reuse strategies can achieve better performances than the other existing reuse schemes for HetNets with RNs.

APPENDIX 1

REUSE PATTERN BANDWIDTH UPDATE & DISTRIBUTED LINK BANDWIDTH UPDATE

A. REUSE PATTERN BANDWIDTH UPDATE

In order to obtain the updated reuse pattern bandwidths in the t -th iteration, we need to solve the optimization problem

in (8a). The Lagrangian for the problem is formulated as

$$\begin{aligned} \mathcal{L}_{\rho_1}(\mathbf{x}, \mathbf{y}^{(t)}, \boldsymbol{\lambda}^{(t)}, \boldsymbol{\mu}, \theta) &= \sum_{u \in \mathcal{M}} \log(W \sum_{n \in \mathcal{N}} \sum_{i \in \mathcal{I}} y_{u,n,i}^{(t)} c_{u,n,i}) \\ &+ \theta \left(\sum_{i \in \mathcal{I}} x_i - 1 \right) + \sum_{n \in \mathcal{N}} \sum_{i \in \mathcal{I}} \lambda_{n,i}^{(t)} (\mathbb{1}\{n \in \mathcal{A}_i\} x_i - \sum_{u \in \mathcal{U}} y_{u,n,i}^{(t)}) \\ &+ \sum_{i \in \mathcal{I}} \mu_i \cdot x_i - \frac{\rho_1}{2} \sum_{n \in \mathcal{N}} \sum_{i \in \mathcal{I}} (\mathbb{1}\{n \in \mathcal{A}_i\} x_i - \sum_{u \in \mathcal{U}} y_{u,n,i}^{(t)})^2, \end{aligned}$$

where θ and $\boldsymbol{\mu} := [\mu_i, \forall i]^T$ are the Lagrange multipliers. Since the gradient of Lagrangian should vanish at the optimal primal and dual points, we have

$$\theta + \mu_i - \rho_1 \sum_{n \in \mathcal{N}} \mathbb{1}\{n \in \mathcal{A}_i\} \left(x_i - \sum_{u \in \mathcal{U}} y_{u,n,i}^{(t)} - \frac{\lambda_{n,i}^{(t)}}{\rho_1} \right) = 0. \quad (25)$$

Then, the optimal reuse profile is derived as

$$x_i^{(t+1)} = \frac{\sum_n \mathbb{1}\{n \in \mathcal{A}_i\} (\lambda_{n,i}^{(t)} + \rho_1 \sum_u y_{u,n,i}^{(t)}) + \theta + \mu_i}{\rho_1 \sum_n \mathbb{1}\{n \in \mathcal{A}_i\}}.$$

Due to the complementary slackness, i.e. $\mu_i \cdot x_i = 0, \forall i \in \mathcal{I}$, the reuse profile can be updated as:

$$x_i^{(t+1)} = \left[\frac{\sum_n \mathbb{1}\{n \in \mathcal{A}_i\} (\lambda_{n,i}^{(t)} + \rho_1 \sum_u y_{u,n,i}^{(t)}) + \theta}{\rho_1 \sum_n \mathbb{1}\{n \in \mathcal{A}_i\}} \right]_0, \quad (26)$$

where θ is chosen to ensure $\sum_{i \in \mathcal{I}} x_i^{(t+1)} = 1$.

B. LINK BANDWIDTH UPDATE

According to the ADMM principle, the augmented Lagrangian for the optimization problem in (10) is given in (27), as shown at the bottom of this page, where $\boldsymbol{\alpha} := [\alpha_k, \forall k]^T$ and $\boldsymbol{\xi} := [\xi_{u,n,i}, \forall u, n, i]^T$ are the Lagrange multipliers, $\boldsymbol{\rho} := [\rho_1, \rho_2, \rho_3]^T$ is the penalty parameter vector. The ADMM solves the problem in (10) by iteratively performing the following four steps in each iteration j :

$$\mathbf{y}^{(j+1)} = \arg \max_{\mathbf{y} \geq \mathbf{0}} \mathcal{L}_{\rho}(\mathbf{y}, \mathbf{z}^{(j)}, \mathbf{x}^{(t+1)}, \boldsymbol{\lambda}^{(t)}, \boldsymbol{\alpha}^{(j)}, \boldsymbol{\xi}^{(j)}), \quad (28)$$

$$\mathbf{z}^{(j+1)} = \arg \max_{\mathbf{z}} \mathcal{L}_{\rho}(\mathbf{y}^{(j+1)}, \mathbf{z}, \mathbf{x}^{(t+1)}, \boldsymbol{\lambda}^{(t)}, \boldsymbol{\alpha}^{(j)}, \boldsymbol{\xi}^{(j)}), \quad (29)$$

$$\alpha_k^{(j+1)} = \alpha_k^{(j)} - \rho_2 (R_{k+M}^{(j+1)} - \tilde{R}_{k+B}^{(j+1)}), \quad \forall k \in \mathcal{K}, \quad (30)$$

$$\xi_{u,n,i}^{(j+1)} = \xi_{u,n,i}^{(j)} - \rho_3 (z_{u,n,i}^{(j+1)} - y_{u,n,i}^{(j+1)}), \quad \forall u, n, i. \quad (31)$$

The Lagrangian for the optimization in (28) is given in (32), as shown at the top of the next page,

$$\begin{aligned} \mathcal{L}_{\rho}(\mathbf{y}, \mathbf{z}, \mathbf{x}^{(t+1)}, \boldsymbol{\lambda}^{(t)}, \boldsymbol{\alpha}, \boldsymbol{\xi}) &= \sum_{u \in \mathcal{M}} \log(W \sum_{i \in \mathcal{I}} \sum_{n \in \mathcal{N}} y_{u,n,i} \cdot c_{u,n,i}) + \sum_{k \in \mathcal{K}} \alpha_k \cdot (R_{k+M} - \tilde{R}_{k+B}) \\ &+ \sum_{n \in \mathcal{N}} \sum_{i \in \mathcal{I}} \lambda_{n,i}^{(t)} \cdot (\mathbb{1}\{n \in \mathcal{A}_i\} \cdot x_i^{(t+1)} - \sum_{u \in \mathcal{U}} z_{u,n,i}) - \frac{\rho_1}{2} \sum_{n \in \mathcal{N}} \sum_{i \in \mathcal{I}} (\mathbb{1}\{n \in \mathcal{A}_i\} \cdot x_i^{(t+1)} - \sum_{u \in \mathcal{U}} z_{u,n,i})^2 \\ &- \frac{\rho_2}{2} \sum_{k \in \mathcal{K}} (R_{k+M} - \tilde{R}_{k+B})^2 + \sum_{u \in \mathcal{U}} \sum_{n \in \mathcal{N}} \sum_{i \in \mathcal{I}} \xi_{u,n,i} (z_{u,n,i} - y_{u,n,i}) - \frac{\rho_3}{2} \sum_{u \in \mathcal{U}} \sum_{n \in \mathcal{N}} \sum_{i \in \mathcal{I}} (z_{u,n,i} - y_{u,n,i})^2, \end{aligned} \quad (27)$$

$$\begin{aligned} \mathcal{L}(\mathbf{y}, \mathbf{z}^{(j)}, \boldsymbol{\lambda}^{(t)}, \boldsymbol{\alpha}^{(j)}, \boldsymbol{\xi}^{(j)}) = & \sum_{u \in \mathcal{M}} \log \left(W \sum_{i \in \mathcal{I}} \sum_{n \in \mathcal{N}} y_{u,n,i} c_{u,n,i} \right) - \frac{\rho_3}{2} \sum_{u \in \mathcal{U}} \sum_{n \in \mathcal{N}} \sum_{i \in \mathcal{I}} (z_{u,n,i}^{(j)} - y_{u,n,i})^2 \\ & + \sum_{k \in \mathcal{K}} \left(\alpha_k R_{k+M} - \frac{\rho_2}{2} (R_{k+M} - \tilde{R}_{k+B}^{(j)})^2 \right) - \sum_{u \in \mathcal{U}} \sum_{n \in \mathcal{N}} \sum_{i \in \mathcal{I}} (\xi_{u,n,i} - \beta_{u,n,i}) y_{u,n,i}, \end{aligned} \quad (32)$$

where $\boldsymbol{\beta} := [\beta_{u,n,i}, \forall u, n, i]^T \succeq \mathbf{0}$ is the Lagrange multiplier. By setting the gradient of the above Lagrangian with respect to \mathbf{y} to zero, the update rule to \mathbf{y} can be obtained as:

$$y_{u,n,i}^{(j+1)} = \left[z_{u,n,i}^{(j)} + \frac{W \cdot c_{u,n,i} \cdot \Gamma_{u,n,i}^{(j)} - \xi_{u,n,i}^{(j)}}{\rho_3} \right]_0, \quad (33)$$

where $\boldsymbol{\beta}$ has been dropped due to the complementary slackness and $\Gamma_{u,n,i}^{(j)}$ is defined as

$$\begin{aligned} \Gamma_{u,n,i}^{(j)} = & \mathbb{1}\{u \leq M\} \cdot \frac{1}{R_u} + \mathbb{1}\{u > M\} \\ & \cdot (\alpha_{u-M}^{(j)} - \rho_2 (R_u - \tilde{R}_{u-M+B}^{(j)})). \end{aligned} \quad (34)$$

The optimization problem in (29) does not have any constraints. By setting the gradient with respect to \mathbf{z} to zero, we can have

$$z_{n,i}^{(j+1)} = (\rho_1 \mathbf{1}\mathbf{1}^T + \rho_3 \mathbf{I})^{-1} \cdot (\mathbf{b}_{n,i}^{(j)} - \mathbb{1}\{n > B\} \cdot W \cdot \boldsymbol{\Lambda}_{n,i}^{(j)}),$$

where

$$\begin{aligned} \mathbf{z}_{n,i}^{(j+1)} & := [z_{1,n,i}^{(j+1)}, \dots, z_{U,n,i}^{(j+1)}]^T, \\ \mathbf{b}_{n,i}^{(j)} & := [b_{1,n,i}^{(j)}, \dots, b_{U,n,i}^{(j)}]^T, \end{aligned}$$

and $\boldsymbol{\Lambda}_{n,i}^{(j)}$ is defined as

$$\boldsymbol{\Lambda}_{n,i}^{(j)} = \mathbf{c}_{n,i} \cdot (\alpha_{n-B}^{(j)} - \rho_2 \cdot (R_{n-B+M}^{(j)} - \tilde{R}_n)), \quad (35)$$

where $\mathbf{c}_{n,i} := [c_{1,n,i}, \dots, c_{U,n,i}]^T$. Note each element of $\mathbf{b}_{n,i}^{(j)}$ is defined as

$$b_{u,n,i}^{(j)} = -\lambda_{n,i}^{(t)} + \rho_1 \mathbb{1}\{n \in \mathcal{A}_i\} x_i^{(t+1)} + \rho_3 y_{u,n,i}^{(j+1)} + \xi_{u,n,i}^{(j)}. \quad (36)$$

APPENDIX 2 PROOF FOR PROPOSITION 1

Assume one particular optimal solution: \mathbf{x} and \mathbf{y} . Define auxiliary variables $\tau_{u,n,i}$ and decompose each element of \mathbf{y} as $y_{u,n,i} = x_i \cdot \tau_{u,n,i}$. The user data rates in (6b) can be rewritten as $R_u = W \sum_i x_i \sum_n \tau_{u,n,i} \cdot c_{u,n,i}, \forall u$. Let $R_u^i := W \sum_n \tau_{u,n,i} \cdot c_{u,n,i}$ be the data rate of the u -th user under reuse pattern- i . The data rates of the M MSs can be expressed as follows:

$$\mathbf{R} = [\mathbf{R}^1, \dots, \mathbf{R}^{2^N-1}] \mathbf{x}, \quad (37)$$

where $\mathbf{R}^i := [R_1^i, \dots, R_M^i]^T$ and $\mathbf{R} := [R_1, \dots, R_M]$. Additionally, according to the constraints in (6d), the optimal reuse profile \mathbf{x} needs to satisfy the following constraints:

$$\sum_i x_i \cdot \left(\sum_n \tau_{k+M,n,i} c_{k+M,n,i} - \sum_u \tau_{u,k+B,i} c_{u,k+B,i} \right) = 0.$$

Define Φ_k^i and Ψ_k^i as $\Phi_k^i := W \sum_n \tau_{k+M,n,i} \cdot c_{k+M,n,i}, \Psi_k^i := W \sum_u \tau_{u,k+B,i} \cdot c_{u,k+B,i}, k \in \mathcal{K}$. The above constraints can be rewritten as the following form:

$$[\Phi^1 - \Psi^1, \dots, \Phi^{2^N-1} - \Psi^{2^N-1}] \mathbf{x} = \mathbf{0}, \quad (38)$$

where $\Phi^i := [\Phi_1^i, \dots, \Phi_K^i]^T$ and $\Psi^i := [\Psi_1^i, \dots, \Psi_K^i]^T$. Combining (37) and (38), the following equation is obtained,

$$[\mathbf{S}^1, \dots, \mathbf{S}^{2^N-1}] \mathbf{x} = \begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix} = \mathbf{S}, \quad (39)$$

where $\mathbf{S}^i := [(\mathbf{R}^i)^T, (\Phi^i)^T - (\Psi^i)^T]^T$. Denote the set of $\{\mathbf{S}^i\}_{i \in \mathcal{I}}$ as \mathcal{P} . According to (6e), the vector \mathbf{S} lies in the convex hull of \mathcal{P} , i.e. $\text{conv}(\mathcal{P})$. Since the dimension of the vector \mathbf{S} is $(M+K)$, from the Carathéodory's Theorem [30], the vector \mathbf{S} must lie in the convex hull of a set of $(M+K+1)$ affinely independent points. Let \mathcal{P}' denote the set of these points and the convex hull of \mathcal{P}' , i.e. $\text{conv}(\mathcal{P}')$, is an $(M+K)$ -simplex.

Note the objective function in (6) is concave with respect to \mathbf{S} . One optimal solution can be found on the face of the $(M+K)$ -simplex due to the Pareto optimality. Therefore, the optimal solution \mathbf{S}^* can be represented by a convex combination of at most $(M+K)$ points in \mathcal{P} . In other words, there exists another reuse profile \mathbf{x}^* satisfying the following conditions:

$$\begin{aligned} \mathbf{S}^* & = [\mathbf{S}^1, \dots, \mathbf{S}^{2^N-1}] \mathbf{x}^* = \begin{bmatrix} \mathbf{R}^* \\ \mathbf{0} \end{bmatrix}, \\ \|\mathbf{x}^*\|_0 & \leq M+K, \quad \sum_{i \in \mathcal{I}} x_i^* = 1, \quad x_i^* \geq 0, \quad \forall i. \end{aligned}$$

APPENDIX 3 SOFT ACTIVE PATTERN IDENTIFICATION

After employing the re-weighted ℓ_1 -norm algorithm, the primal-dual updates in (20a) and (20c) in the t -th iteration are modified as:

$$\mathbf{x}^{(t+1)} = \arg \max_{\mathbf{x} \in \mathcal{X}} \tilde{\mathcal{L}}_{\rho_1}(\mathbf{x}, \mathbf{y}^{(t)}, \boldsymbol{\lambda}^{(t)}, \kappa^{(t)}); \quad (40a)$$

$$\kappa^{(t+1)} = \kappa^{(t)} - \delta_\kappa \cdot \left(D - \sum_{i \in \mathcal{I}} w_i^{(t+1)} \cdot x_i^{(t+1)} \right), \quad (40b)$$

where δ_κ is the step size in updating κ . Similar to the derivations in (25) - (26) in Appendix 1-A, we can have the following closed-form update for (40a):

$$\begin{aligned} x_i^{(t+1)} & = \left[\frac{\sum_n \mathbb{1}\{n \in \mathcal{A}_i\} (\lambda_{n,i}^{(t)} + \rho_1 \sum_u y_{u,n,i}^{(t)}) - \kappa^{(t)} w_i^{(t)} + \theta}{\rho_1 \sum_n \mathbb{1}\{n \in \mathcal{A}_i\}} \right]_0, \end{aligned} \quad (41)$$

where θ is chosen such that $\sum_{i \in \mathcal{I}} x_i^{(t+1)} = 1$.

The update for the Lagrange multiplier κ does not have a closed-form solution. It is thus updated with the gradient descent method. As discussed in [30], the size of the step size δ_κ affects the converge behavior towards the optimum.

REFERENCES

- [1] A. Damnjanovic et al., "A survey on 3GPP heterogeneous networks," *IEEE Wireless Commun.*, vol. 18, no. 3, pp. 10–21, Jun. 2011.
- [2] A. Ghosh et al., "Heterogeneous cellular networks: From theory to practice," *IEEE Commun. Mag.*, vol. 50, no. 6, pp. 54–64, Jun. 2012.
- [3] R. Pabst et al., "Relay-based deployment concepts for wireless and mobile broadband radio," *IEEE Commun. Mag.*, vol. 42, no. 9, pp. 80–89, Sep. 2004.
- [4] C. Hoymann, W. Chen, J. Montojo, A. Golitschek, C. Koutsimanis, and X. Shen, "Relaying operation in 3GPP LTE: Challenges and solutions," *IEEE Commun. Mag.*, vol. 50, no. 2, pp. 156–162, Feb. 2012.
- [5] H. Zhang, C. Jiang, N. C. Beaulieu, X. Chu, X. Wen, and M. Tao, "Resource allocation in spectrum-sharing OFDMA femtocells with heterogeneous services," *IEEE Trans. Commun.*, vol. 62, no. 7, pp. 2366–2377, Jul. 2014.
- [6] P. Xue, P. Gong, J. H. Park, D. Park, and D. K. Kim, "Radio resource management with proportional rate constraint in the heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 11, no. 3, pp. 1066–1075, Mar. 2012.
- [7] A. Agustin and J. Vidal, "Amplify-and-forward cooperation under interference-limited spatial reuse of the relay slot," *IEEE Trans. Wireless Commun.*, vol. 7, no. 5, pp. 1952–1962, May 2008.
- [8] N. Saquib, E. Hossain, and D. I. Kim, "Fractional frequency reuse for interference management in LTE-advanced hetnets," *IEEE Wireless Commun.*, vol. 20, no. 2, pp. 113–122, Apr. 2013.
- [9] B. Zhuang, D. Guo, and M. L. Honig, "Traffic-driven spectrum allocation in heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 10, pp. 2027–2038, Oct. 2015.
- [10] B. Zhuang, D. Guo, and M. L. Honig, "Energy-efficient cell activation, user association, and spectrum allocation in heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 823–831, Apr. 2016.
- [11] Q. Kuang, W. Utschick, and A. Dotzler, "Optimal joint user association and multi-pattern resource allocation in heterogeneous networks," *IEEE Trans. Signal Process.*, vol. 64, no. 13, pp. 3388–3401, Jul. 2016.
- [12] X. Luo, "Delay-oriented QoS-aware user association and resource allocation in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1809–1822, Mar. 2017.
- [13] R. Etkin, A. Parekh, and D. Tse, "Spectrum sharing for unlicensed bands," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 3, pp. 517–528, Apr. 2007.
- [14] M. Guizani, B. Khalfi, M. Ben Ghorbel, and B. Hamdaoui, "Large-scale cognitive cellular systems: Resource management overview," *IEEE Commun. Mag.*, vol. 53, no. 5, pp. 44–51, May 2015.
- [15] S. Bhattarai, J.-M. Park, B. Gao, K. Bian, and W. Lehr, "An overview of dynamic spectrum sharing: Ongoing initiatives, challenges, and a roadmap for future research," *IEEE Trans. Cogn. Commun. Netw.*, vol. 2, no. 2, pp. 110–128, Jun. 2016.
- [16] M. Peng, C. Wang, J. Li, H. Xiang, and V. Lau, "Recent advances in underlay heterogeneous networks: Interference control, resource allocation, and self-organization," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 700–729, Apr. 2015.
- [17] T. O. Olwal, K. Djouani, and A. M. Kurien, "A survey of resource management toward 5G radio access networks," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 1656–1686, 3rd Quart., 2016.
- [18] Y. L. Lee, T. C. Chuah, J. Loo, and A. Vinel, "Recent advances in radio resource management for heterogeneous LTE/LTE-A networks," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 4, pp. 2142–2180, 4th Quart., 2014.
- [19] M. Salem et al., "An overview of radio resource management in relay-enhanced OFDMA-based networks," *IEEE Commun. Survey Tuts.*, vol. 12, no. 3, pp. 422–438, 3rd Quart., 2010.
- [20] W. Shim, Y. Han, and S. Kim, "Fairness-aware resource allocation in a cooperative OFDMA uplink system," *IEEE Trans. Veh. Technol.*, vol. 59, no. 2, pp. 932–939, Feb. 2010.
- [21] J. Lee, H. Wang, W. Seo, and D. Hong, "QoS-guaranteed transmission mode selection for efficient resource utilization in multi-hop cellular networks," *IEEE Trans. Wireless Commun.*, vol. 7, no. 10, pp. 3697–3701, Oct. 2008.
- [22] O. Oyman, "Opportunistic scheduling and spectrum reuse in relay-based cellular networks," *IEEE Trans. Wireless Commun.*, vol. 9, no. 3, pp. 1074–1085, Mar. 2010.
- [23] Q. Li, R. Q. Hu, Y. Qian, and G. Wu, "Intracell cooperation and resource allocation in a heterogeneous network with relays," *IEEE Trans. Veh. Technol.*, vol. 62, no. 4, pp. 1770–1784, May 2013.
- [24] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Nov. 2010.
- [25] E. J. Candès, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *J. Fourier Anal. Appl.*, vol. 14, nos. 5–6, pp. 877–905, 2008.
- [26] Z. Zhu, S. Jin, Y. Yang, H. Hu, and X. Luo, "Time reusing in D2D-enabled cooperative networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 3185–3200, Feb. 2018.
- [27] A. Sabharwal, P. Schniter, D. Guo, D. W. Bliss, S. Rangarajan, and R. Wichman, "In-band full-duplex wireless: Challenges and opportunities," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 9, pp. 1637–1652, Sep. 2014.
- [28] M. Grant and S. Boyd. (Sep. 2013). *CVX: MATLAB Software for Disciplined Convex Programming, Version 2.0 Beta*. [Online]. Available: <http://cvxr.com/cvx>
- [29] M. Johansson and L. Xiao, "Cross-layer optimization of wireless networks using nonlinear column generation," *IEEE Trans. Wireless Commun.*, vol. 5, no. 2, pp. 435–445, Feb. 2006.
- [30] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms*, 3rd ed. Hoboken, NJ, USA: Wiley, 2006.
- [31] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. B, Methodol.*, vol. 58, no. 1, pp. 267–288, 1996.
- [32] B.-G. Choi, I. Doh, and M. Y. Chung, "Radio resource management scheme for relieving interference to MUEs in relay-based cellular networks," *IEEE Trans. Veh. Technol.*, vol. 64, no. 7, pp. 3018–3029, Jul. 2015.
- [33] W.-C. Liao, M. Hong, Y.-F. Liu, and Z.-Q. Luo, "Base station activation and linear transceiver design for optimal resource management in heterogeneous networks," *IEEE Trans. Signal Process.*, vol. 62, no. 15, pp. 3939–3952, Aug. 2014.
- [34] D. Fooladivanda and C. Rosenberg, "Joint resource allocation and user association for heterogeneous wireless cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, pp. 248–257, Jan. 2013.



SHENGDA JIN received the B.S. degree from Zhejiang University, Hangzhou, China, in 2015, and the M.Sc. degree in electrical engineering from the School of Information Science and Technology, ShanghaiTech University. His research interests include resource allocation in wireless communication, fog computing, and signal processing on graph.



ZHAOWEI ZHU received the B.S. degree from the University of Electronic Science and Technology of China, Chengdu, China, in 2016. He is currently pursuing the M.S. degree with ShanghaiTech University, Shanghai, China, under the guidance of Prof. X. Luo. His research interests lie in the distributed machine learning and signal processing methods for mobile networks, fog computing, and signal processing on graph.

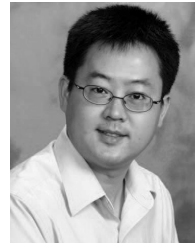


YUECHEN WU is currently pursuing the B.Eng. degree in electrical engineering with the School of Information Science and Technology, ShanghaiTech University, under the guidance of Prof. X. Luo. His interests include distributed computing, convex optimization, machine learning, and signal processing.



SADIQ ALI received the B.Sc. degree in electrical engineering from the University of Engineering and Technology, Peshawar, Pakistan, in 2004, the M.Sc. degree in digital communication systems and technology from the Department of Signals and Systems, Chalmers University of Technology, Gothenburg, Sweden, in 2008, and the Ph.D. degree in signal processing from the Universitat Autònoma de Barcelona (UAB), Spain, in 2014. He was as a Research Assistant at the

Microelectronics and Electronics Systems Department, UAB, from 2008 to 2009. From 2014 to 2015, he was a Post-Doctoral Research Fellow at Qatar University. He is currently a Faculty Member and Researcher at the University of Engineering and Technology, Peshawar. His research interests include distributed detection and estimation, wireless sensor networks, time-series and time-frequency analysis, array processing, and communication systems.



XILIANG LUO (S'03–M'06–SM'18) received the B.Sc. degree in physics from Peking University, Beijing, China, in 2001, and the M.Sc. and Ph.D. degrees in electrical engineering from the University of Minnesota, Minneapolis, MN, USA, in 2003 and 2006, respectively.

After finishing his Ph.D. studies, he joined Qualcomm Research and carried out cutting edge research at different posts as a Senior Engineer in 2006, a Staff Engineer in 2010, and then a Senior Staff Engineer in 2013, where he was involved in the system designs, analyses, and standardization of 4G LTE. He was the Designer of various enhancements to Qualcomm's current LTE solutions and led the designs of Qualcomm's next generation LTE modem for heterogeneous networks from initial concept to final completion. Since 2014, he has been with the School of Information Science and Technology, ShanghaiTech University, Shanghai, China, as an Associate Professor. He has authored or co-authored over 70 research papers in top journals and conferences. He is the co-inventor of over 70 U.S. and international patents, the majority of which have been adopted into current LTE and LTE-Advanced standards. He is currently the Co-Director of the Shanghai Institute of Fog Computing Technology. His current research interests include signal processing, communications, and information theory. In particular, he is interested in researches combining information theory and signal processing theory that can shape and guide the designs of next generation data and information processing networks. In 2017, he received the Excellent Paper Award from the IEEE ICUFN. He is also serving as an Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS.

• • •