A Novel Approach for QoS-Aware Joint User Association, Resource Block and Discrete Power Allocation in HetNets

Hamza Umit Sokun, *Student Member, IEEE*, Ramy H. Gohary, *Senior Member, IEEE*, and Halim Yanikomeroglu, *Fellow, IEEE*

***Abstract*— We consider joint optimization of user-to-base-station (BS) association, and time-frequency resource block (RB) and power allocation in heterogeneous networks (HetNets). The objective is to develop a design: 1) that maximizes the number of users accommodated in the network while satisfying their quality of service demands and 2) that minimizes usage of the resources required to meet these demands. We investigate two novel instances of HetNets with opportunistic RB-reuse. In the first instance, user-to-BS associations and power allocations can be time-shared, and the RBs can be reused during the signaling interval. For this instance, it is shown that the design problem can be approximated by a problem that yields tight convex upper and lower bounds on the objective. In contrast, the second instance represents a case in which the RBs can be reused, but the user-to-BS associations and power allocations are not time-shared, and hence, fixed throughout the signaling interval. The latter case gives rise to a combinatorial optimization problem, which we provide an approximate solution for by using a polynomial-complexity two-phase approach based on semidefinite relaxation with randomization.**

***Index Terms*— HetNet, convex optimization, semidefinite relaxation, gaussian randomization, time-sharing, opportunistic RB-reuse.**

I. Introduction

HETEROGENEOUS networks (HetNets) are wireless networks that incorporate low-power pico and femto base stations (BSs) in addition to conventional macro BSs. The low cost and flexibility of deploying such networks can not only be used to increase their capacity, but also to extend their coverage and to provide better user experience. To realize the benefits of HetNets and to attain desired network performance, the interplay between various network functionalities must be taken into consideration. For instance, the way in which users are associated to the BSs and the time-frequency resource blocks (RBs) available to the system can have a fundamental impact on the number of users that can be accommodated. Likewise, the way in which the BSs allocate their available power across RBs can have a significant impact on the amount of physical resources required for the network to attain particular performance objectives.

For the users to be associated to the BSs in a resource-efficient manner, several considerations must be taken into account. These include the channel conditions between the users and the BSs, the load condition of each BS, and the quality-of-service (QoS) demanded by each user.[1] A conventional approach for associating users to BSs is the one in which each user is associated to the BS with the maximum signal-to-interference-plus-noise ratio (max-SINR). This approach is simple, but can lead to highly imbalanced BS loading and does not guarantee users' QoS. In particular, this methodology can result in severely overloading the high-power macro BSs while leaving the low-power pico and femto BSs largely underutilized. Such load imbalance does not take advantage of the pico and femto BSs in supporting more users in the network.

In addition to user-to-BS association, the number of users that can be accommodated in the network is affected by the way in which the RBs are allocated to each user [2], [3]. For instance, granting a user the right to utilize its preferred RB can result in depriving other users of the opportunity to communicate altogether. Hence, it can be seen that accommodating more users with QoS constraints in the network requires a mechanism that takes into consideration not only the local network conditions of individual users, but rather the global conditions of all users. In a complementary fashion, it is desirable for the design to utilize minimal physical resources to accommodate a given network load. For instance, even if the number of users is fixed and the optimal user-to-BS associations are known *a priori*, the available RBs must be allocated to the users in a manner that avoids excessive consumption of valuable physical resources.

Efficient utilization of resources is coupled with the way in which power is allocated across the available RBs. For practical considerations, transmit powers usually assume

Manuscript received August 26, 2016; revised February 11, 2017 and June 12, 2017; accepted August 27, 2017. Date of publication September 15, 2017; date of current version November 9, 2017. This work was supported in part by Huawei Canada Co., Ltd., and in part by the Ontario Ministry of Economic Development and Innovation's Ontario Research Fund-Research Excellence Program. This paper was presented at the IEEE Vehicular Technology Conference (VTC-Fall), Boston, MA, USA, September 2015. The associate editor coordinating the review of this paper and approving it for publication was L. Dai. *(Corresponding author: Hamza Umit Sokun.)*

The authors are with the Department of Systems and Computer Engineering, Carleton University, Ottawa, ON K1S 5B6, Canada (e-mail: husokun@sce.carleton.ca; gohary@sce.carleton.ca; halim@sce.carleton.ca).

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

Digital Object Identifier 10.1109/TWC.2017.2751467

[1] Similar to [1], herein we assume that QoS is measured by the data rate delivered to the users.

discrete, rather than continuous, levels. For instance, networks based on the Long Term Evolution (LTE) standards can only support discrete power levels in the downlink [4], [5]. Restricting the power allocations to be discrete simplifies the hardware design and reduces the cost of practical transmitters. To see the effect of the available power on the allocation of RBs, we note that a set of RBs can be allocated to a particular user only if the power available at the BS to transmit on those RBs suffices to meet the user's demanded QoS. Hence, allocating power across RBs is coupled by the way in which the RBs are assigned to the users, and by the way in which the users are assigned to the BSs. Sequential consideration of these functionalities simplifies design, at the expense of reducing the user capacity and exhausting valuable network resources. To avoid these drawbacks it is essential to jointly optimize the user-to-BS associations and the RBs and the power allocations to ensure that the maximum number of QoS-constrained users can be accommodated with the least possible physical resources.

In this paper, two instances of downlink HetNets with opportunistic RB-reuse are investigated. In the first instance the user-to-BS associations, and the RBs and power level allocations are time-shared [6], which implies that the corresponding assignment variables assume continuous values. Using this, we develop convex formulations for tight upper and lower bounds on the design objective. In the second instance, no time-sharing is considered and the user-to-BS associations, and the RBs and power level allocations, once determined, are restricted to be fixed during the entire signalling interval. In this case the assignment variables assume binary values and the resulting optimization problem is non-convex. In both instances, the formulations automatically invoke opportunistic RB-reuse. In other words, for each RB, they allow each BS to either use the RB exclusively at any given instance or to share it concurrently with another BS, i.e., they do not enforce the BSs to reuse RBs. Moreover, in both instances each user can be associated by one or multiple BSs as in [7]–[9]. The case with time-sharing generalizes the one without time-sharing. Hence, it offers the potential of achieving significant performance advantages, however at the expense of increasing signalling overhead. This is especially true in networks with time-varying channels [6]. In such situations, it might be more practical to restrict the RB schedules to be binary, thereby implying that each RB is assigned to a particular user for the entire signalling interval [2], [10]. Generally speaking, user-to-BS association typically takes place at a time scale larger than the one in which the resources are allocated. However, jointly optimizing the user-to-BS associations and the resource allocations can be used at system initialization and at periodic, yet relatively wide, intervals thereafter. This will provide the system with the optimal signalling parameters at those periodic instants. In between times, subproblems for optimizing power and RB allocations with fixed user-to-BS associations can be readily derived from the formulations provided in our work. Furthermore, joint optimization of the user-to-BS association and power and RB allocations, even though not always feasible, provides a benchmark for the maximum number of users that can be supported in the network.

A. Related Work

Methods for determining 'good' user-to-BS associations have been considered in the literature, mostly in isolation from other network design aspects. For instance, associating users to the BSs with the highest SINR, as in the max-SINR technique, was the conventional method up to LTE release-8 [1]. Despite being arguably the simplest approach, it does not take the effect of loading into account and can therefore result in severe load imbalance across the system. A modified approach that attempts to alleviate this drawback was proposed in [11]. In this approach, a user is associated with the BSs that would yield the maximum expected throughput rather than the SNR. Since the expected throughput depends on the signal strength and the population of users served by the BS, this heuristic approach provides better load balancing in comparison to the max-SINR one. Another heuristic technique for user-to-BS association is the one based on range expansion (RE) [12]. In that approach, the traffic from macro BSs is offloaded to low-power BSs by adding a positive bias to their measured SINRs. An advantage of this approach is that it does not depend on the relative location of users and BSs. However, a major weakness of it is the lack of guidance for choosing the biasing factor. Algorithms for determining user-to-BS associations by finding approximate solutions to relaxed network utility maximization problems using dual pricing methods were proposed in [7] and [13]. Those methods do not account for the QoS requirements. Other approaches that invoke game-theoretic techniques for generating user-to-BS associations have been considered in [14] and [15]. The techniques that underlie those approaches do not necessarily converge and the associations that they yield can be highly suboptimal. In [16], semidefinite relaxation (SDR) technique were employed in a HetNet scenario, which is restricted to have one macro and one pico BS per cell and the design objective was to maximize the total sum rate irrespective of the QoSs demanded by the users. Despite the potential benefits offered by these approaches, the fact that they do not incorporate other key design aspects, such as RB and power allocations, usually limits the extent of their practical effectiveness.

To increase their effectiveness, the user-to-BS association techniques have been considered in conjunction with power allocation ones in [17]–[19]. Such techniques were developed for various HetNet design objectives, but with fixed RB allocations, which usually facilitates formulating the design problem. For instance, in [17] the design problem corresponding to maximizing the number of users in the system considering user-to-BS associations and continuous power allocations was formulated for the case in which the users have prescribed QoS demands. The resulting problem is non-convex, but was solved in [17] using a technique based on a branch-and-bound type approach called Bender's decomposition. This technique is efficient in many practical scenarios, but its worst case complexity is non-polynomial. A somewhat related problem, is the one considered in [18]. Therein, the user-to-BS associations and the continuous power allocations that maximize the minimum SINR of the users of the downlink of a HetNet were investigated, but without QoS guarantees. It was shown in [18] that this problem is NP-hard unless the number of BSs is equal to the number of users. In the latter case, the problem has polynomial complexity. Other instances of jointly optimizing user-to-BS associations and power allocations in HetNets were considered in, e.g., [19]. Finding the global solution in such instances constitutes a formidable task, but local solutions could be obtained efficiently using iterative techniques.

The efficiency with which the physical resources are utilized can be further enhanced by jointly considering user-to-BS associations along with both RB and power allocations. Although no results seem to be available for the practical case of discrete power allocations, the case with continuous power allocations seem to have attracted attention. For instance, such power allocations are jointly optimized with user-to-BS associations and RB allocations using a greedy approach in [20] to maximize a proportional fairness metric of the rates delivered to the HetNet users. Similarly, user-to-BS associations, RB and continuous power allocations are jointly optimized in [21] using a relaxed formulation and an iterative approach that aims at maximizing the throughput of the HetNet. Another approach that considers optimizing these aspects jointly is developed in [22] using, the so-called, Gibbs sampler. This approach aims at minimizing the system-wide delay experienced by the users, but without considering their QoS demands.

B. Contribution and Comparison With Related Work

Having provided some background on the designs that consider the optimization of user-to-BS associations, and RBs and power allocations either separately or jointly, we now provide a brief account of our contributions. Specifically, the contributions of the paper are the following: 1) We introduce a novel framework for jointly optimizing the user-to-BS associations, and the RBs and power allocations in HetNets with potentially disparate power budgets and spectral resources. 2) We consider two cases: with and without time-sharing. For both instances, we develop optimization frameworks and polynomial-complexity algorithms that maximizes the number of users accommodated in the network and, at the same time, minimizes the network resources consumed to accommodate those users. The designs proposed herein for the two instances possess the following distinguishing features in comparison with related work.

- The design objective considered herein is to jointly maximize the number of network users while ensuring minimal consumption of the physical resources. Hence, this objective relates directly to the revenue of the system operator and the satisfaction of a larger set of users [17]. This is in contrast with other objectives that are commonly considered in the literature. For instance, sum-rate maximization may result in a small number of users being served by the network and may therefore impact the revenue of the service provider.
- The transmit power levels considered herein are constrained to assume discrete, rather than continuous, values. This feature renders the algorithms developed herein amenable to practical systems. Although most existing resource allocation schemes consider continuous transmit power, applying those schemes in practical systems can be problematic; rounding does not guarantee optimality. Hence, it can be seen that our work makes a step towards bridging the divide between theoretical-designs and practical implementation of resource allocation algorithms. We show that, when time-sharing is considered, imposing the restriction on the power levels to be discrete does not affect the convexity of the formulations for upper and lower bounds. However, when time-sharing is not considered, the restriction on the power levels to be discrete usually results in difficult-to-solve non-convex formulations. This difficulty is alleviated here by invoking the SDR technique which usually generates close-to-optimal performance by invoking a careful relaxation on the rank of the matrix-value optimization variable. This technique enables to perform joint optimization for HetNets with a relatively large number of BSs, RBs, users, and power levels. Such an optimization would be computationally prohibitive to perform with exhaustive search.
- The proposed algorithms in the time-sharing and the no time-sharing cases allow the RBs to be reused opportunistically. Hence, these algorithms offer valuable design flexibility and enable higher spectral efficiencies to be achieved.

II. System Model

In this section we describe the downlink of a two-tier HetNet, in which tier-1 consists of one or more macro BSs and tier-2 consists of one or more pico BSs. The main difference between these BSs is the available power budgets; macro BSs have power budgets that are manifold of the power budgets of their pico counterparts. Subsequently, the area of the cell covered by a pico BS is usually much smaller than that covered by a macro BS.

It is desirable in such a HetNet scenario to optimize the network parameters in order to maximize the number of users supported by the network, while minimizing the amount of consumed resources. To achieve this goal, we consider a centralized design whereby a central entity that is aware of the network parameters controls the association of users to BSs, the allocation of RBs among users, and the allocation of available power budget across RBs. In our work, we consider a setup similar to the one considered in [7] and [13], wherein the BSs, macro or pico, share a common pool of RBs. This setup offers design flexibility and avoids the penalty of partitioning the available frequency band into non-overlapping sections. Now, depending on the power budget of each BS and the channel to prospective users, it may be beneficial for one BS to serve one user on a particular RB, but not on another one. The availability of the RB depends on whether it is being used to serve other users, and whether RB-reuse is allowed. In a complementary fashion, it can be seen that an RB may be suitable to serve a particular user from one BS, but not from another BS, depending on the aforementioned factors. Hence, it can be seen that, whether RB-reuse is allowed or not, RB assignment and BS association are tightly coupled; a change in one aspect can/will readily affect the other aspect.

To expose the relationship between the network parameters and the design objective, we denote the set of all BSs, including macro and pico ones, by $\mathcal{B} = \{1, \ldots, B\}$ and the set of all users by $\mathcal{U} = \{1, \ldots, U\}$. The wide range of services, e.g., voice and non-voice applications, requested by those users render their QoS demands rather disparate. We denote the QoS demanded by user u by Q_u, $u \in \mathcal{U}$, and it will be measured in megabits per second (Mbps). To satisfy these demands, the service provider allocates a different number of RBs to each user; higher QoS requirement implies higher usage of RBs. In addition, the usage of RBs depends on the channel conditions observed by the user; weaker channels imply higher usage of RBs. For notational convenience, we will denote the set of RBs available to all the BSs in the system by $\mathcal{S} = \{1, \ldots, S\}$.

For practical considerations, it is desirable for each BS to use a finite set of discrete power levels [5]. We will denote the set of power levels utilized by the b-th BSs by $\mathcal{P}_b = \{p_b^1, \ldots, p_b^L\}$, where $L = |\mathcal{P}_b|$ is the cardinality of $\mathcal{P}_b$, $b \in \mathcal{B}$. Note that, because we are considering a scenario with a heterogeneous infrastructure, the set $\mathcal{P}_b$ may vary across BSs. For instance, this set may depend on the total amount of transmit power available at each BS.

To complete the characterization of the considered HetNet, we will denote the channel gain between the b-th BS and the u-th user on the s-th RB by h_{ub}^s, $u \in \mathcal{U}$, $b \in \mathcal{B}$, $s \in \mathcal{S}$. These gains account for path-loss, log-normal shadowing, and fading and will be assumed to be random and fixed during the signalling interval. We will use W_0 to denote the bandwidth of each RB, and N_0 to denote the power spectral density of the additive white Gaussian noise at the receiver.

In the scenario considered herein, a central entity is assumed to have access to accurate channel estimates and power budgets. The central entity optimizes the user-to-BS associations, and the RB and power allocations and sends its decisions to the users and the BSs. Two cases can be considered, viz., without and with RB-reuse. In the first case, the central entity opts for design and processing simplicity by avoiding RB-reuse and assigning each RB exclusively to one user throughout [23]. In contrast, in the second case, the central entity opts for higher spectral efficiency by allowing RB-reuse and concurrently assigning each RB to potentially multiple users. The second case subsumes the first one. In other words, allowing RB-reuse, in the sense of the forthcoming formulations, encompasses the case with no RB-reuse. Hence, we will focus on the case with RB-reuse, and allude to the other case as necessary. To avoid overly complicated exposition, we will restrict attention on the case in which at most two users within the HetNet can share the same RB at any given instant. Allowing each RB to be shared concurrently by multiple users results in contaminating the users' signal with interference from neighbouring BSs. In this case, the power of the signal received by the u-th user, $u \in \mathcal{U}$, on the s-th RB, $s \in \mathcal{S}$, from the b-th BS, $b \in \mathcal{B}$, when it uses the ℓ-th power level, $\ell \in \mathcal{P}_b$, is given by $p_b^\ell |h_{bu}^s|^2$. The interference signal that this user observes from the k-th BS, $k \in \mathcal{B} \setminus b$, when it uses the n-th power level, $n \in \mathcal{P}_k$, is given by $p_k^n |h_{ku}^s|^2$. Subsequently, the SINR of the signal can be expressed as

$$\gamma_{bu}^{s\ell kn} = \frac{p_b^\ell |h_{bu}^s|^2}{W_0 N_0 + p_k^n |h_{ku}^s|^2}, \tag{1}$$

and the corresponding rate is given by

$$R_{bu}^{s\ell kn} = W_0 \log_2(1 + \gamma_{bu}^{s\ell kn}) \quad \text{[bits per second]}. \tag{2}$$

For notational convenience, we will use $\gamma_{bu}^{s\ell 00} = \frac{p_b^\ell |h_{bu}^s|^2}{W_0 N_0}$ to represent the signal-to-noise ratio (SNR) when the s-th RB is not reused. To proceed, we note that $\gamma_{bu}^{s\ell kn}$ and $R_{bu}^{s\ell kn}$ are determined by the ordered sextet (b, u, s, ℓ, k, n), that is, a given sextet (b, u, s, ℓ, k, n) corresponds to values of $\gamma_{bu}^{s\ell kn}$ and $R_{bu}^{s\ell kn}$ that can be determined *a priori*. We will use this observation in the prospective formulation to distill the optimization of user-to-BS associations, and RBs and power allocations to the optimization of sextets as will be elucidated in the next section.

III. Objective and System Constraints

In this section, we will use the sextet (b, u, s, ℓ, k, n) to describe the design objective and the system constraints. We will show that the problem of associating users to BSs and allocating RBs and power levels amounts to determining an optimal set of sextets from among all admissible ones. To determine admissible sextets, we introduce an indicator variable $y_{bu}^{s\ell kn}$ for each sextet (b, u, s, ℓ, k, n), $b \in \mathcal{B}$, $u \in \mathcal{U}$, $s \in \mathcal{S}$, $\ell \in \mathcal{P}_b$, $k \in \{0\} \cup \mathcal{B} \setminus b$, and $n \in \mathcal{P}_k$. When time-sharing is considered, $y_{bu}^{s\ell kn}$ represents the fraction of time during which the sextet (b, u, s, ℓ, k, n) is used for transmission, and hence, $y_{bu}^{s\ell kn} \in [0, 1]$. In contrast, when time-sharing is not considered, $y_{bu}^{s\ell kn}$ is 1 if the sextet (b, u, s, ℓ, k, n) is used for transmission and is zero otherwise, i.e., $y_{bu}^{s\ell kn} \in \{0, 1\}$.

A. System Constraints

1) BS Power Budget Constraints: In a HetNet, the total transmission power budget may vary significantly between macro and pico BSs. We will denote such a budget by $P_b^{\max}$, $b \in B$. Hence, the total power transmitted by the b-th BS to all the users in the network must satisfy

$$\sum_{u \in \mathcal{U}} \sum_{s \in \mathcal{S}} \sum_{\ell \in \mathcal{P}_b} \Big(y_{bu}^{s\ell 00} + \sum_{k \in \mathcal{B} \setminus b} \sum_{n \in \mathcal{P}_k} y_{bu}^{s\ell kn} \Big) p_b^\ell \leq P_b^{\max}, \quad b \in \mathcal{B}. \tag{3}$$

2) QoS Constraints: The objective of the design developed herein is to maximize the number of users. Hence, it must ensure that the users' QoS demands are met. Otherwise, infinitely many users can be trivially accommodated, i.e., with vanishing QoSs. To avoid such a possibility, the QoS metric is taken to be the data rate that can be reliably communicated to each user, e.g., [1]. Using (2), the maximum aggregate data rate that can be reliably communicated to user $u \in \mathcal{U}$, can be expressed as $\sum_{b \in \mathcal{B}} \sum_{s \in \mathcal{S}} \sum_{\ell \in \mathcal{P}_b} \big(y_{bu}^{s\ell 00} R_{bu}^{s\ell 00} + \sum_{k \in \mathcal{B} \setminus b} \sum_{n \in \mathcal{P}_k} y_{bu}^{s\ell kn} R_{bu}^{s\ell kn} \big)$. Hence, ensuring that the QoS requirement of user u is satisfied is equivalent to ensuring that

$$\sum_{b \in \mathcal{B}} \sum_{s \in \mathcal{S}} \sum_{\ell \in \mathcal{P}_b} \Big(y_{bu}^{s\ell 00} R_{bu}^{s\ell 00} + \sum_{k \in \mathcal{B} \setminus b} \sum_{n \in \mathcal{P}_k} y_{bu}^{s\ell kn} R_{bu}^{s\ell kn} \Big) \geq Q_u, \quad u \in \mathcal{U}. \tag{4}$$

This constraint must be satisfied only for those users accommodated in the network. Otherwise, it must be excluded from the formulation. However, since the design objective is to maximize the number of users, the users that will be actually accommodated in the network are not known *a priori*. To circumvent this difficulty, in Section IV-A, we will provide a mathematically convenient means for characterizing this conditional constraint automatically.

3) RB Allocation Constraints: To derive a constraint on the RB allocation, let us first consider the case without RB-reuse. In this case, restricting each RB to be used at most once at any given time instant can be enforced by constraining the summation of $\{y_{bu}^{s\ell 00}\}$ over all BSs, users and power levels to be at most equal to one. Next, we consider the case with RB-reuse in which two BSs use the same RB. In this case, the summation of $\{y_{bu}^{s\ell kn}\}$ over all BSs, users, power levels and combinations of interfering BSs with their power levels to be at most equal to two. Since our formulation allows each RB to be reused opportunistically, i.e., it can be used exclusively by one BS during a fraction of the signalling interval and RB-reused during the remainder of the signalling interval, the following constraint must be satisfied.

$$\sum_{b\in\mathcal{B}}\sum_{u\in\mathcal{U}}\sum_{\ell\in\mathcal{P}_b} y_{bu}^{s\ell 00}+\frac{1}{2}\sum_{b\in\mathcal{B}}\sum_{u\in\mathcal{U}}\sum_{\ell\in\mathcal{P}}\sum_{k\in\mathcal{B}\setminus b}\sum_{n\in\mathcal{P}_k} y_{bu}^{s\ell kn}\leq 1,\quad s\in\mathcal{S}. \tag{5}$$

4) RB-Reuse Constraints: Since in our formulation at most two BSs can concurrently use the same RB, and each BS can use one power level at any time instant, we have:

$$y_{bu}^{s\ell 00}+\sum_{k\in\mathcal{B}\setminus b}\sum_{n\in\mathcal{P}_k} y_{bu}^{s\ell kn}\leq 1,\quad b\in\mathcal{B},\ u\in\mathcal{U},\ s\in\mathcal{S},\ \ell\in\mathcal{P}_b. \tag{6}$$

This constraint enables RB-reuse to be invoked opportunistically; the case with no RB-reuse corresponds to $y_{bu}^{s\ell kn}=0$, for every $b\in\mathcal{B}$, $\ell\in\mathcal{P}_b$, $u\in\mathcal{U}$, $k\in\mathcal{B}\setminus b$ and $n\in\mathcal{P}_k$.

5) Interference-Coupling Constraints: Let us consider a time-sharing case in which a particular time fraction $y_{bu}^{s\ell kn}$ is strictly greater than zero (For the corresponding case without time-sharing, $y_{bu}^{s\ell kn}=1$.). Such $y_{bu}^{s\ell kn}$ implies that the k-th BS is transmitting a signal with power level n for at least a time fraction of $y_{bu}^{s\ell kn}$, $b\in\mathcal{B}$, $\ell\in\mathcal{P}_b$, $u\in\mathcal{U}$, $k\in\mathcal{B}\setminus b$ and $n\in\mathcal{P}_k$. This implies that $y_{k\tilde{u}}^{snb\ell}$ must be strictly greater than zero for at least one $\tilde{u}\in\mathcal{U}\backslash u$. But, since a BS can transmit to multiple users at different time instants, it follows that $y_{bu}^{s\ell kn}$ must satisfy the following constraint:

$$y_{bu}^{s\ell kn}\leq\sum_{\tilde{u}\in\mathcal{U}\backslash u} y_{k\tilde{u}}^{snb\ell},\quad b\in\mathcal{B},\ u\in\mathcal{U},\ s\in\mathcal{S},\ \ell\in\mathcal{P}_b,\quad k\in\mathcal{B}\setminus b, n\in\mathcal{P}_k. \tag{7}$$

We make two observations. First, for the case with RB-reuse and no time-sharing when $y_{bu}^{s\ell kn}=1$, the constraint in (7) ensures that $y_{k\tilde{u}}^{snb\ell}=1$ for exactly one user $\tilde{u}\in\mathcal{U}\setminus u$. Second, (7) is relevant when $k\neq 0$, and is trivial otherwise, i.e., for RBs that are not reused.

6) User Association Constraints: In our framework, a BS cannot simultaneously use multiple power levels nor broadcast to multiple users on the same RB. This yields the following constraint:

$$\sum_{\ell\in\mathcal{P}_b}\sum_{u\in\mathcal{U}}\left(y_{bu}^{s\ell 00}+\sum_{k\in\mathcal{B}\setminus b}\sum_{n\in\mathcal{P}_k} y_{bu}^{s\ell kn}\right)\leq 1,\quad b\in\mathcal{B},\ s\in\mathcal{S}. \tag{8}$$

7) BS Association Constraints: At any time instant and RB, each user can be served by at most one BS with one power level. This yields the following constraint:

$$\sum_{b\in\mathcal{B}}\sum_{\ell\in\mathcal{P}_b}\left(y_{bu}^{s\ell 00}+\sum_{k\in\mathcal{B}\setminus b}\sum_{n\in\mathcal{P}_k} y_{bu}^{s\ell kn}\right)\leq 1,\quad u\in\mathcal{U},\ s\in\mathcal{S}. \tag{9}$$

B. Design Objective

The design objective considered in this paper is composed of two components. The first characterizes the number of users accommodated by the system. Maximizing this component directly contributes to maximizing revenue and utility. The second component characterizes RB usage and ought to be minimized provided that the accommodated users' QoS constraints are satisfied. We will begin by providing the mathematical description of the second component.

1) Aggregate RB Usage: The aggregate usage of RBs assigned to the u-th user can be expressed as $\sum_{b\in\mathcal{B}}\sum_{s\in\mathcal{S}}\sum_{\ell\in\mathcal{P}_b}\left(y_{bu}^{s\ell 00}+\sum_{k\in\mathcal{B}\setminus b}\sum_{n\in\mathcal{P}_k} y_{bu}^{s\ell kn}\right)$, $u\in\mathcal{U}$. This usage has slightly different meanings in the cases with and without time-sharing. To see that, consider the case in which there is only one RB. When time-sharing is considered, the summation of $\{y_{bu}^{s\ell kn}\}$ corresponds to the normalized total time over which this RB is used. Conversely, when time-sharing is not considered, this summation serves as an indicator whether this RB is used or not. Hence, the aggregate usage of RBs in the network can be expressed as $\sum_{u\in\mathcal{U}}\sum_{b\in\mathcal{B}}\sum_{s\in\mathcal{S}}\sum_{\ell\in\mathcal{P}_b}\left(y_{bu}^{s\ell 00}+\sum_{k\in\mathcal{B}\setminus b}\sum_{n\in\mathcal{P}_k} y_{bu}^{s\ell kn}\right)$. Minimizing this summation directly reduces the amount of resources exhausted by the system, and indirectly contributes to increasing revenue.

2) Total Number of Accommodated Users: The number of users accommodated is indirectly related to RB usage, and the number of BSs and power levels. To see that, we recall that in the current system each user may be served by multiple BSs over multiple RBs with different power levels at every time instant. Hence, a user $u\in\mathcal{U}$ is guaranteed to be served whenever at least one of the assignment variables in the set $\{y_{bu}^{s\ell kn}\}$, $s\in\mathcal{S}$, $\ell\in\mathcal{P}_b$, $b\in\mathcal{B}$, $k\in\{0\}\cup\mathcal{B}\backslash b$, $n\in\mathcal{P}_k$, is strictly greater than zero. But since $\{y_{bu}^{s\ell kn}\}$ are restricted to be non-negative, it can be readily seen that a user $u\in\mathcal{U}$ is guaranteed to be served if and only if $\sum_{b\in\mathcal{B}}\sum_{s\in\mathcal{S}}\sum_{\ell\in\mathcal{P}_b}\left(y_{bu}^{s\ell 00}R_{bu}^{s\ell 00}+\sum_{k\in\mathcal{B}\setminus b}\sum_{n\in\mathcal{P}_k} y_{bu}^{s\ell kn}R_{bu}^{s\ell kn}\right)>0$. In other words, a user $u\in\mathcal{U}$ is not accommodated in the system if and only if $\sum_{b\in\mathcal{B}}\sum_{s\in\mathcal{S}}\sum_{\ell\in\mathcal{P}_b}\left(y_{bu}^{s\ell 00}R_{bu}^{s\ell 00}+\sum_{k\in\mathcal{B}\setminus b}\sum_{n\in\mathcal{P}_k} y_{bu}^{s\ell kn}R_{bu}^{s\ell kn}\right)=0$. Hence, to indicate whether a particular user $u\in\mathcal{U}$ is accommodated in the network or not, we define the indicator function $\imath(x)$ such that $\imath(x)=1$ for $x>0$ and $\imath(x)=0$ for $x=0$. To use $\imath(x)$ as a user accommodation indicator, the variable x is replaced with the normalized[2] aggregate rate of the u-th user, $\sum_{b\in\mathcal{B}}\sum_{s\in\mathcal{S}}\sum_{\ell\in\mathcal{P}_b}\frac{1}{Q_u}\big(y_{bu}^{s\ell 00}R_{bu}^{s\ell 00}+\sum_{k\in\mathcal{B}\backslash b}\sum_{n\in\mathcal{P}_k}y_{bu}^{s\ell kn}R_{bu}^{s\ell kn}\big)$. Unfortunately, the indicator function is not smooth and hence cannot be readily incorporated in the optimization framework. To circumvent this difficulty, we use an alternate representation. We express it as

$$\imath(x)=\lim_{\sigma\to\infty}1-e^{-\sigma x}. \tag{10}$$

This representation implies that if σ is chosen to be sufficiently large, $1-e^{-\sigma x}$ will provide a smooth approximation of $\imath(x)$. A proper choice of σ depends on whether time-sharing is considered or not. In particular, when time-sharing is considered, we will show that the choice of a good value of σ is straightforward, whereas when time-sharing is not considered, a judicious choice of this value must be made. We will elaborate on the choice of the value of σ for the latter case in Section V-B.

3) A Composite Objective: Our design objective is composed of the linear combination of two quantities: 1) the number of accommodated users, which ought to be maximized to increase revenue, and 2) the aggregate RB usage, which ought to be minimized to efficiently utilize the network resources. To synthesize this objective, we will weight each of its components by an appropriate scalar $\rho\in[0,1]$. Using the scalar ρ, we define our design objective to be maximizing $\rho\sum_{u\in\mathcal{U}}(1-e^{-\frac{\sigma}{Q_u}\sum_{b\in\mathcal{B}}\sum_{s\in\mathcal{S}}\sum_{\ell\in\mathcal{P}_b}(y_{bu}^{s\ell 00}R_{bu}^{s\ell 00}+\sum_{k\in\mathcal{B}\backslash b}\sum_{n\in\mathcal{P}_k}y_{bu}^{s\ell kn}R_{bu}^{s\ell kn})})-(1-\rho)\sum_{u\in\mathcal{U}}\sum_{b\in\mathcal{B}}\sum_{s\in\mathcal{S}}\sum_{\ell\in\mathcal{P}_b}(\sum_{k\in\mathcal{B}\backslash b}\sum_{n\in\mathcal{P}_k}y_{bu}^{s\ell kn}+y_{bu}^{s\ell 00})$. Increasing ρ from 0 to 1 in this objective shifts the emphasis of the optimization from minimizing the usage of the resources used in the network to maximizing the number of users accommodated therein. As such, ρ parameterizes a family of objectives, each of which corresponds to a number of accommodated users and a level of aggregate RB usage. In order to ensure that the number of users accommodated in the network is maximized, the optimal choice of the value of ρ is stated in the following proposition.

Proposition 1: Any value of $\rho\in\left(\frac{2S}{1+2S},1\right)$ ensures maximizing the number of users accommodated in the network.

Proof: See Appendix A. ■

IV. Optimization Based Formulation With and Without Time-Sharing

A. Network Instance With Time-Sharing

We consider a network instance with time-sharing, where the indicator variables $\{y_{bu}^{s\ell kn}\}$ are continuous on the interval [0, 1]. Using a value of ρ within the range suggested in Proposition 1, we can now provide a mathematical formulation for the design that jointly optimizes the user-to-BS associations, the RBs and the power allocations. Before we do that, we provide a mathematically convenient means for characterizing the QoS constraint in (4). To do that, we will use the approximation of $\imath(\cdot)$ in Section III-B. In particular, the constraint in (4) is equivalent to multiplying the right hand side of the inequality therein with the indicator function used in the design objective with σ sufficiently large. Using this characterization, the joint design problem can be cast in the form given at the top of next page in (11). In this formulation, (11c) and (11d), as shown at the top of the next page, are equivalent to (4) when $\sigma\to\infty$. In other words, similar to (4), for large σ, (11c) and (11d) ensure that the QoS constraints are enforced only for the users accommodated in the network and are immaterial otherwise; unaccommodated users cannot impose QoS constraints. When $\sigma\to\infty$, t_u in (11d) indicates whether the u-th user $u\in\mathcal{U}$ is accommodated in the network or not. If accommodated, $t_u=0$, otherwise $t_u=1$.

Before proceeding with our formulation, we note that it is usually possible for the optimization of time-shares to be cast in a convex form because time-sharing of two feasible solutions is equivalent to finding a point in their convex hull. However, the objective of the optimization at hand is not convex because of the (discrete) number of users to be maximized. This discreteness is captured by (11d) with large σ. Hence, although time-sharing results in a convex feasible set, the design problem remains non-convex because of the discrete component of the objective.

The accuracy of the approximation in the objective, (11a), and the constraints in (11c) and (11d) is controlled by σ; the value of σ is generally not critical, and choosing σ to be large, e.g., $\sigma=100$, ensures that each summand in the first term of the objective and the right hand side of (11c) is essentially equal to one whenever the summation in the exponent is greater than zero. In other words, choosing σ in (11d) to be large ensures that whenever the aggregate user rate normalized to its QoS is bounded away from zero, $t_u\approx 0$ and (11c) is enforced; when the normalized aggregate user rate is close to zero $t_u\approx 1$ and (11c) is trivial.

From the formulation in (11), it can be seen that all constraints are linear except the non-affine equality constraint in (11d), which renders the formulation non-convex, and hence difficult to solve. To circumvent this difficulty, we relax the constraint (11d) to be an inequality in the form

$$t_u\ \geq\ e^{-\frac{\sigma}{Q_u}\sum_{b\in\mathcal{B}}\sum_{s\in\mathcal{S}}\sum_{\ell\in\mathcal{P}}\left(y_{bu}^{s\ell 00}R_{bu}^{s\ell 00}+\sum_{k\in\mathcal{B}\backslash b}\sum_{n\in\mathcal{P}}y_{bu}^{s\ell kn}R_{bu}^{s\ell kn}\right)},\quad u\in\mathcal{U}. \tag{12}$$

This constraint together with the objective imply that when the normalized rate of the u-th user is zero, t_u is equal to one and (11c) holds with equality. However, when the normalized rate of the u-th user is greater than zero, the new constraint in (12) may be inactive, and t_u can take on a value between 0 and 1. In this case, the fact that the objective is monotonically increasing in $(1-t_u)$, $u\in\mathcal{U}$, implies that (11c) will be active. In other words, for every user, $u\in\mathcal{U}$, at least one of the two inequality constraints in (11c), and (12) will be satisfied with equality. More specifically, let us define $\alpha_u\ \triangleq\ \frac{\sum_{b\in\mathcal{B}}\sum_{s\in\mathcal{S}}\sum_{\ell\in\mathcal{P}}\left(y_{bu}^{s\ell 00}R_{bu}^{s\ell 00}+\sum_{k\in\mathcal{B}\backslash b}\sum_{n\in\mathcal{P}}y_{bu}^{s\ell kn}R_{bu}^{s\ell kn}\right)}{Q_u}$. We have the following three operating regions:

1) When $\alpha_u=0$, there is no resource allocated to user u and its rate is zero. In this case, both (11c), and (12) hold with equality and the relaxation is tight.

[2]Normalization by the QoS of individual users is introduced for numerical convenience.

$$\max_{\mathbf{y},\ \mathbf{t}} \quad \rho \sum_{u \in \mathcal{U}} (1 - t_u) - (1 - \rho) \sum_{s \in \mathcal{S}} \sum_{b \in \mathcal{B}} \sum_{u \in \mathcal{U}} \sum_{\ell \in \mathcal{P}_b} \Big(y_{bu}^{s\ell 00} + \sum_{k \in \mathcal{B} \setminus b} \sum_{n \in \mathcal{P}_k} y_{bu}^{s\ell kn} \Big), \tag{11a}$$

$$\text{subject to} \quad \sum_{u \in \mathcal{U}} \sum_{s \in \mathcal{S}} \sum_{\ell \in \mathcal{P}_b} \Big(y_{bu}^{s\ell 00} + \sum_{k \in \mathcal{B} \setminus b} \sum_{n \in \mathcal{P}_k} y_{bu}^{s\ell kn} \Big) p_b^{\ell} \leq P_b^{\max}, \quad b \in \mathcal{B}, \tag{11b}$$

$$\sum_{b \in \mathcal{B}} \sum_{s \in \mathcal{S}} \sum_{\ell \in \mathcal{P}_b} \Big(y_{bu}^{s\ell 00} R_{bu}^{s\ell 00} + \sum_{k \in \mathcal{B} \setminus b} \sum_{n \in \mathcal{P}_k} y_{bu}^{s\ell kn} R_{bu}^{s\ell kn} \Big) \geq Q_u (1 - t_u), \quad u \in \mathcal{U}, \tag{11c}$$

$$t_u = e^{-\frac{\sigma}{Q_u} \sum_{b \in \mathcal{B}} \sum_{s \in \mathcal{S}} \sum_{\ell \in \mathcal{P}_b} \left(y_{bu}^{s\ell 00} R_{bu}^{s\ell 00} + \sum_{k \in \mathcal{B} \setminus b} \sum_{n \in \mathcal{P}_k} y_{bu}^{s\ell kn} R_{bu}^{s\ell kn} \right)}, \quad u \in \mathcal{U}, \tag{11d}$$

$$\sum_{b \in \mathcal{B}} \sum_{u \in \mathcal{U}} \sum_{\ell \in \mathcal{P}_b} y_{bu}^{s\ell 00} + \frac{1}{2} \sum_{b \in \mathcal{B}} \sum_{u \in \mathcal{U}} \sum_{\ell \in \mathcal{P}_k} \sum_{k \in \mathcal{B} \setminus b} \sum_{n \in \mathcal{P}} y_{bu}^{s\ell kn} \leq 1, \quad s \in \mathcal{S}, \tag{11e}$$

$$y_{bu}^{s\ell 00} + \sum_{k \in \mathcal{B} \setminus b} \sum_{n \in \mathcal{P}_k} y_{bu}^{s\ell kn} \leq 1, \quad b \in \mathcal{B},\ u \in \mathcal{U},\ s \in \mathcal{S},\ \ell \in \mathcal{P}_b, \tag{11f}$$

$$y_{bu}^{s\ell kn} \leq \sum_{\tilde{u} \in \mathcal{U} \setminus u} y_{k\tilde{u}}^{snb\ell}, \quad b \in \mathcal{B},\ u \in \mathcal{U},\ s \in \mathcal{S},\ \ell \in \mathcal{P}_b,\ k \in \mathcal{B} \setminus b,\ n \in \mathcal{P}_k, \tag{11g}$$

$$\sum_{\ell \in \mathcal{P}_b} \sum_{u \in \mathcal{U}} \Big(y_{bu}^{s\ell 00} + \sum_{k \in \mathcal{B} \setminus b} \sum_{n \in \mathcal{P}_k} y_{bu}^{s\ell kn} \Big) \leq 1, \quad b \in \mathcal{B},\ s \in \mathcal{S}, \tag{11h}$$

$$\sum_{b \in \mathcal{B}} \sum_{\ell \in \mathcal{P}_b} \Big(y_{bu}^{s\ell 00} + \sum_{k \in \mathcal{B} \setminus b} \sum_{n \in \mathcal{P}_k} y_{bu}^{s\ell kn} \Big) \leq 1, \quad u \in \mathcal{U},\ s \in \mathcal{S}, \tag{11i}$$

$$t_u \geq 0, \quad u \in \mathcal{U}, \tag{11j}$$

$$y_{bu}^{s\ell kn} \in [0, 1], \quad b \in \mathcal{B},\ u \in \mathcal{U},\ s \in \mathcal{S},\ \ell \in \mathcal{P}_b,\ k \in \{0\} \cup \mathcal{B} \setminus b,\ n \in \mathcal{P}_k. \tag{11k}$$

2) When $\alpha_u \geq 1$, the QoS requirement of user u is achieved. In this case, both (11c), and (12) hold with equality and the relaxation is tight.
3) When $\alpha_u \in (0, 1)$, the QoS requirement of user u is neither low enough to guarantee its support in the network nor high enough to be definitely denied access to the network. In this case, either (11c), or (12) holds with equality and the other one holds with strict inequality, and the relaxation is not tight.

Now we will show how (12) will be used to obtain upper and lower bounds on the number of supported users. Using (12), the relaxed version of the problem in (11) can be expressed as in (13), given at the top of the next page. The formulation in (13) comprises an efficiently solvable convex problem; objective and constraints are convex. To see how (13) can be used to generate upper and lower bounds on the objective, we consider the following possibilities:

- If (12) holds with equality for every $u \in \mathcal{U}$, then (13) is solved optimally, and the cardinality of the set of u's for which the corresponding $t_u = 0$ is equal to the maximum number of users that can be supported in the network.
- If (12) does not hold with equality for a set of users in, say $\tilde{\mathcal{U}} \subset \mathcal{U}$, we consider the cardinality of this set, i.e., $|\tilde{\mathcal{U}}|$. We have two cases: either $|\tilde{\mathcal{U}}| = 1$, or $|\tilde{\mathcal{U}}| > 1$.
  - If $|\tilde{\mathcal{U}}| = 1$, then t_u assumes a fractional value for exactly one user, i.e., $t_u \in (0, 1)$. But, since one component of the objective is to minimize $\sum_{u \in \mathcal{U}} t_u$, it can be readily seen that not allowing the user with fractional t_u to join the network is optimal, and the cardinality of the set of u's for which the corresponding $t_u = 0$ is equal to the maximum number of users that can be supported in the network.
  - If $|\tilde{\mathcal{U}}| > 1$, then the cardinality of the set of u's for which the corresponding $t_u = 0$ provides a lower bound on the maximum number of users that can be supported. In this case, an upper bound on the maximum number of users can be obtained by assuming that the users in $\tilde{\mathcal{U}}$ can also be incorporated in the network.

In particular, suppose that the maximum number of users than can be accommodated in the network is given by U^*, i.e., $U^* = \sum_{u \in \mathcal{U}} (1 - t_u^*)$, where t_u^* is generated by (11). To obtain an upper bound on U^*, we note that when $\{t_u\}$ generated by (13) assume fractional values, the first component of the objective in (13), $\sum_{u \in \mathcal{U}} (1 - t_u)$, also assumes a fractional value and an upper bound on U^* is given by $\left\lfloor \sum_{u \in \mathcal{U}} (1 - t_u) \right\rfloor$. To obtain a lower bound, we round each non-zero value of t_u to 1. This yields that a lower bound on the number of users that can be accommodated in the system is given by $\sum_{u \in \mathcal{U}} \lfloor (1 - t_u) \rfloor$. In other words, the maximum number of users than can be accommodated in the network, U^*, satisfies the following inequalities:

$$\textstyle\sum_{u \in \mathcal{U}} \lfloor (1 - t_u) \rfloor \leq U^* \leq \lfloor \sum_{u \in \mathcal{U}} (1 - t_u) \rfloor. \tag{14}$$

We conclude by noting that, even though the power levels in (13) are constrained to be discrete, the fact that these levels are determined by indicator functions over a continuous interval enables us to develop a convex framework for generating tight bounds on the objective. This framework represents an instance in which design complexity is traded for dimensionality.

$$\max_{\mathbf{y},\ \mathbf{t}} \quad \rho \sum_{u\in\mathcal{U}}(1-t_u) - (1-\rho)\sum_{s\in\mathcal{S}}\sum_{b\in\mathcal{B}}\sum_{u\in\mathcal{U}}\sum_{\ell\in\mathcal{P}_b}\Big(y_{bu}^{s\ell 00} + \sum_{k\in\mathcal{B}\backslash b}\sum_{n\in\mathcal{P}_k} y_{bu}^{s\ell kn}\Big), \tag{13a}$$

$$\text{subject to (11b), (11c), (11e)–(11k) and (12).} \tag{13b}$$

B. Network Instance Without Time-Sharing

In contrast with the time-sharing-based formulation, herein we consider an instance of a network without time-sharing. This framework yields a formulation analogous to that in (11) but with the continuous constraint in (11k) replaced with the binary constraint $y_{bu}^{s\ell kn} \in \{0, 1\}$, $\forall b, u, s, \ell, k, n$. We have the optimization problem given in (15), as shown at the top of the next page. The formulation in (15) is non-convex because of (15c) and (11d) in (15b). Similar to (13), an upper bound on (15) can be obtained by relaxing (11d), which yields the formulation given in (16), as shown at the top of the next page. Apart from the binary constraint in (16c), the formulation in (16) constitutes a convex optimization problem. However, because of the binary constraint, this problem represents an instance of a non-convex mixed integer nonlinear program (MINLP). The complexity of algorithms for solving such MINLP problems is generally high, rendering their applicability in the current scenario questionable. To alleviate this difficulty, we will use a polynomial-complexity technique based on SDR with randomization [24]. However, unlike the time-sharing case, the effectiveness of the SDR-based technique depends on the choice of σ in (12), cf. Section V-B below.

V. Solution Approaches

A. Network Instance With Time-Sharing

The formulation for the optimization problem in (13) is convex; the objective is linear, and the constraints in (13b) are linear except the constraint (12), which is convex. Hence, it can be solved reliably and efficiently in polynomial time using, e.g., interior-point methods (IPM) [25].

1) Computational Complexity Analysis: The problem in (13) is solved only once, and the computational complexity of solving it with IPM can be obtained using the analysis pertaining to self-concordant functions. For such functions the total number of Newton steps is proportional to the number of inequality constraints. Fortunately, when the logarithmic barrier function is invoked, the problem in (13) possesses the self-concordance property. We have the following result.

Proposition 2: The computational complexity of solving the problem in (13) can be bounded by $\mathcal{O}(n^{3.5})$, where $n = B + S + 3U + BS + US + BUSL + BUSL(2(B-1)L+1)$.

Proof: See Appendix B. ■

It can be seen from this result that the computational complexity of solving (13) is polynomial in B, S, U, and L, which implies that it can be efficiently solved for relatively large networks.

2) Impact of RB-Reuse Order on Computational Complexity: Increasing the order of RB-reuse can offer a substantial increase in throughput and utility. However, this comes at a large computational cost. In particular, the number of optimization variables increases quickly with the order of reuse. This renders solving the formulation corresponding to the general case of arbitrary RB-reuse excessively time consuming, thereby reducing its practical significance. To see that, we note that the number of variables corresponding to an RB-reuse of order q can be shown to be given by $BUSL\sum_{r=0}^{q-1}\frac{(B-1)!}{(B-r-1)!}L^r$. This number grows at least exponentially with q, and hence results in a computationally prohibitive formulation except for small q. Notwithstanding the magnitude of the number of variables, our formulation is readily extensible to general RB-reuse orders, but with more optimization variables.

B. Network Instance Without Time-Sharing

For room considerations, in this section we will provide explicit development of the SDR-based technique for the case with no time-sharing and no RB-reuse. Carrying over this development to the case with RB-reuse is straightforward, but consumes unaffordable room. The formulation for the case with no RB-reuse can be readily derived from (16) by dropping all the $y_{bu}^{s\ell kn}$ terms for which $k \neq 0$. For notational convenience, in this case, the $k = n = 0$ in the superscript can be dropped, and since the constraints in (11h) and (11i) are automatically incorporated in (11e), they can be dropped as well.

Similar to its RB-reuse counterpart, the no RB-reuse variant of (16) represents an instance of an MINLP problem, which is computationally infeasible to solve optimally for all, but a relatively small class of HetNets with a limited number of users, BSs, RBs, and power levels. To obtain resource-efficient solutions for larger networks, we will use an SDR-based technique, which, unlike techniques based on the branch-and-bound approach, has polynomial complexity. This renders it suitable for solving (16) and its no RB-reuse variant for more practical network scenarios. In comparison with other polynomial-complexity algorithms, when constraints and objective are quadratic, the SDR-based technique has a guaranteed approximation accuracy. For instance, for the (NP-hard) maximum-cut problem, the SDR-based technique is guaranteed to yield at least 87.56% of the optimal objective value [26].

To cast the formulation for the case with no RB-reuse in a form that is more amenable to the proposed SDR-based technique, we will express the objective and the constraints using vector notation. Towards that end, let Y be a 4-dimensional tensor with the entries given by $y_{bu}^{s\ell}$, $b \in \mathcal{B}$, $u \in \mathcal{U}$, $s \in \mathcal{S}$, $\ell \in \mathcal{P}_b$. For ease of representation, we express this tensor in the form of a $BS \times UL$ block-partitioned matrix. In particular, Y is written as a matrix of $B \times U$

$$\max_{\mathbf{y},\ \mathbf{t}} \quad \rho \sum_{u\in\mathcal{U}}(1-t_u) - (1-\rho)\sum_{s\in\mathcal{S}}\sum_{b\in\mathcal{B}}\sum_{u\in\mathcal{U}}\sum_{\ell\in\mathcal{P}_b}\Big(y_{bu}^{s\ell 00} + \sum_{k\in\mathcal{B}\setminus b}\sum_{n\in\mathcal{P}_k} y_{bu}^{s\ell kn}\Big), \tag{15a}$$

$$\text{subject to (11b)–(11j),} \tag{15b}$$

$$y_{bu}^{s\ell kn} \in \{0,1\},\quad b\in\mathcal{B},\ u\in\mathcal{U},\ s\in\mathcal{S},\ \ell\in\mathcal{P}_b,\ k\in\{0\}\cup\mathcal{B}\setminus b,\ n\in\mathcal{P}_k. \tag{15c}$$

$$\max_{\mathbf{y},\ \mathbf{t}} \quad \rho \sum_{u\in\mathcal{U}}(1-t_u) - (1-\rho)\sum_{s\in\mathcal{S}}\sum_{b\in\mathcal{B}}\sum_{u\in\mathcal{U}}\sum_{\ell\in\mathcal{P}_b}\Big(y_{bu}^{s\ell 00} + \sum_{k\in\mathcal{B}\setminus b}\sum_{n\in\mathcal{P}_k} y_{bu}^{s\ell kn}\Big), \tag{16a}$$

$$\text{subject to (11b), (11c), (11e)–(11j) and (12),} \tag{16b}$$

$$y_{bu}^{s\ell kn} \in \{0,1\},\quad b\in\mathcal{B},\ u\in\mathcal{U},\ s\in\mathcal{S},\ \ell\in\mathcal{P}_b,\ k\in\{0\}\cup\mathcal{B}\setminus b,\ n\in\mathcal{P}_k. \tag{16c}$$

blocks, $Y = \begin{bmatrix} Y_{11} & \dots & Y_{1U} \\ \vdots & \ddots & \vdots \\ Y_{B1} & \dots & Y_{BU} \end{bmatrix}$, where $Y_{bu} = \begin{bmatrix} y_{bu}^{11} & \dots & y_{bu}^{1L} \\ \vdots & \ddots & \vdots \\ y_{bu}^{S1} & \dots & y_{bu}^{SL} \end{bmatrix}$, $b \in \mathcal{B}$, $u \in \mathcal{U}$. To facilitate subsequent analysis, we introduce three additional 4-dimensional tensors: D_s, E_u, and F_b. These tensors will serve as indicators for users, BSs, RBs and power levels, respectively. Similar to Y, these tensors are expressed in the form of a $BS \times UL$ block-partitioned matrix with $B \times U$ blocks, each with $S \times L$ entries. We define these tensors as follows: the tensor D_s can be expressed as $D_s = \begin{bmatrix} D_{s_{11}} & \cdots & D_{s_{1U}} \\ \vdots & \ddots & \vdots \\ D_{s_{B1}} & \cdots & D_{s_{BU}} \end{bmatrix}$, $s \in \mathcal{S}$, where, for all $i \in \mathcal{B}$ and $j \in \mathcal{U}$, $D_{s_{ij}} = \boldsymbol{i}_s \mathbf{1}_L^T$, where $\boldsymbol{i}_s$ is the s-th column of the $S \times S$ identity matrix $\boldsymbol{I}_S$ and $\mathbf{1}_L$ is the L-dimensional all-one vector. The tensor E_u can be expressed as $E_u = \begin{bmatrix} E_{u_{11}} & \cdots & E_{u_{1U}} \\ \vdots & \ddots & \vdots \\ E_{u_{B1}} & \cdots & E_{u_{BU}} \end{bmatrix}$, $u \in \mathcal{U}$, where $E_{u_{ij}} = \begin{bmatrix} R_{iu}^{11} & \dots & R_{iu}^{1L} \\ \vdots & \ddots & \vdots \\ R_{iu}^{S1} & \dots & R_{iu}^{SL} \end{bmatrix}$ for $i \in \mathcal{B}$ and $j \in \mathcal{U}$, when $j = u$, and $E_{u_{ij}} = \mathbf{0}_{S\times L}$ when $j \neq u$. Finally, the tensor F_b can be written as $F_b = \begin{bmatrix} F_{b_{11}} & \cdots & F_{b_{1U}} \\ \vdots & \ddots & \vdots \\ F_{b_{B1}} & \cdots & F_{b_{BU}} \end{bmatrix}$, $b \in \mathcal{B}$, where $F_{b_{ij}} = \begin{bmatrix} p_b^1 & \dots & p_b^L \\ \vdots & \ddots & \vdots \\ p_b^1 & \dots & p_b^L \end{bmatrix}$ for $i \in \mathcal{B}$ and $j \in \mathcal{U}$, when $i = b$, and $F_{b_{ij}} = \mathbf{0}_{S\times L}$ when $i \neq b$.

Using vec$(\cdot)$ to denote the operator that stacks the columns of a matrix on top of each other, we make the following definitions: $\mathbf{y} \triangleq \mathrm{vec}(Y^T)$, $\mathbf{d}_s \triangleq \mathrm{vec}(D_s^T)$, $\mathbf{e}_u \triangleq \mathrm{vec}(E_u^T)$ and $\mathbf{f}_b \triangleq \mathrm{vec}(F_b^T)$. For ease of exposition, in the forthcoming formulations we will drop the subscript indicating the dimension of the all-one and the all-zero vectors and matrices.

To cast the optimization problem (16) in a form amenable to SDR, we introduce the vector $\boldsymbol{\beta} \in \{-1, 1\}^{BUSL}$, $\boldsymbol{\beta} = 2\mathbf{y} - \mathbf{1}$, where $\boldsymbol{y} \in \{0, 1\}^{BUSL}$, i.e.,

$$\mathbf{y} = \frac{1}{2}(\boldsymbol{\beta} + \mathbf{1}). \tag{17}$$

Using (17), the formulation in (16) can be cast as

$$\max_{\mathbf{y},\ t_u} \quad \rho\sum_{u\in\mathcal{U}}(1-t_u) - (1-\rho)\frac{\mathbf{1}^T}{2}(\boldsymbol{\beta}+\mathbf{1}), \tag{18a}$$

$$\text{subject to } \frac{\mathbf{d}_s^T}{2}(\boldsymbol{\beta}+\mathbf{1}) \le 1, \quad s\in\mathcal{S}, \tag{18b}$$

$$\frac{\mathbf{f}_b^T}{2}(\boldsymbol{\beta}+\mathbf{1}) \le P_b^{\max}, \quad b\in\mathcal{B}, \tag{18c}$$

$$\frac{\mathbf{e}_u^T}{2}(\boldsymbol{\beta}+\mathbf{1}) \ge Q_u(1-t_u), \quad u\in\mathcal{U}, \tag{18d}$$

$$t_u \ge e^{-\sigma\frac{\mathbf{e}_u^T(\boldsymbol{\beta}+\mathbf{1})}{2Q_u}}, \quad u\in\mathcal{U}, \tag{18e}$$

$$\boldsymbol{\beta} \in \{-1,1\}^{BUSL}. \tag{18f}$$

We note that this problem is not convex because of the constraint in (18f). We also note that, for the solution of the relaxed problem in (18) to be optimal for the original problem in (15), the constraints in (18e) must be satisfied with equality $\forall u \in \mathcal{U}$. We will now, show how the SDR technique can be used to provide an approximate solution for (18).

To use the SDR technique, the optimization variables in (18) are constrained to be in the cone of symmetric positive semidefinite (PSD) matrices [24]. Towards that end, we define the following vectors in $\mathbb{R}^{BUSL+1}$, $\hat{\mathbf{d}}_s \triangleq [\mathbf{d}_s^T\ \ \mathbf{d}_s^T\mathbf{1}]^T$, $s \in \mathcal{S}$, $\hat{\mathbf{f}}_b \triangleq [\mathbf{f}_b^T\ \ \mathbf{f}_b^T\mathbf{1}]^T$, $b \in \mathcal{B}$, $\hat{\mathbf{e}}_u \triangleq [\mathbf{e}_u^T\ \ \mathbf{e}_u^T\mathbf{1}]^T$, $u \in \mathcal{U}$, $\hat{\mathbf{1}} \triangleq [\mathbf{1}^T\ \ \mathbf{1}^T\mathbf{1}]^T$, $\hat{\boldsymbol{\beta}} \triangleq [\boldsymbol{\beta}\ \ 1]^T$ and $\hat{\mathbf{k}} \triangleq [\mathbf{0}^T\ \ 1]^T$, and the PSD matrices $\boldsymbol{\Phi} \in \mathbb{R}^{BUSL\times BUSL}$ and $\boldsymbol{\Psi} \in \mathbb{R}^{(BUSL+1)\times(BUSL+1)}$ to be $\boldsymbol{\Phi} \triangleq \boldsymbol{\beta}\boldsymbol{\beta}^T$ and $\boldsymbol{\Psi} = \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^T$, i.e., $\boldsymbol{\Psi} = \begin{bmatrix} \boldsymbol{\Phi} & \boldsymbol{\beta} \\ \boldsymbol{\beta}^T & 1 \end{bmatrix}$. Finally, we define the $(BUSL+1)\times(BUSL+1)$ matrices $\mathbf{A}_1 \triangleq \hat{\mathbf{k}}\hat{\mathbf{1}}^T$, $\mathbf{A}_{\hat{\mathbf{d}}_s} \triangleq \hat{\mathbf{k}}\hat{\mathbf{d}}_s^T$, $\mathbf{A}_{\hat{\mathbf{f}}_b} \triangleq \hat{\mathbf{k}}\hat{\mathbf{f}}_b^T$ and $\mathbf{A}_{\hat{\mathbf{e}}_u} \triangleq \hat{\mathbf{k}}\hat{\mathbf{e}}_u^T$. Using these definitions, it can be readily verified that (16) is equivalent to

$$\max_{\boldsymbol{\Psi},\ t_u} \quad \rho\sum_{u\in\mathcal{U}}(1-t_u) - \frac{(1-\rho)}{2}\mathrm{Tr}(\mathbf{A}_1\boldsymbol{\Psi}), \tag{19a}$$

$$\text{subject to } \frac{1}{2}\mathrm{Tr}(\mathbf{A}_{\hat{\mathbf{d}}_s}\boldsymbol{\Psi}) \le 1, \quad s\in\mathcal{S}, \tag{19b}$$

$$\frac{1}{2}\mathrm{Tr}(\mathbf{A}_{\hat{\mathbf{f}}_b}\boldsymbol{\Psi}) \le P_b^{\max}, \quad b\in\mathcal{B}, \tag{19c}$$

$$\frac{1}{2}\mathrm{Tr}(\mathbf{A}_{\hat{\mathbf{e}}_u}\boldsymbol{\Psi}) \ge Q_u(1-t_u), \quad u\in\mathcal{U}, \tag{19d}$$

$$\mathrm{diag}(\boldsymbol{\Psi}) = 1, \tag{19e}$$

$$t_u \ge e^{-\frac{\sigma}{2Q_u}\mathrm{Tr}(\mathbf{A}_{\hat{\mathbf{e}}_u}\boldsymbol{\Psi})}, \quad u\in\mathcal{U}, \tag{19f}$$

$$\boldsymbol{\Psi} \succeq \mathbf{0}, \tag{19g}$$

$$\mathrm{rank}(\boldsymbol{\Psi}) = 1. \tag{19h}$$

Similar to the case in Section IV-A, it can be seen that a user $u \in \mathcal{U}$ is not accommodated in the system if and only if $\mathrm{Tr}(\mathbf{A}_{\hat{\mathbf{e}}_u}\mathbf{\Psi}) = 0$. In other words, a user $u \in \mathcal{U}$ is guaranteed to be served if and only if $\mathrm{Tr}(\mathbf{A}_{\hat{\mathbf{e}}_u}\mathbf{\Psi}) > 0$. A sufficient condition for this to happen is that $t_u = 0$, cf. (19f). In other words, the QoS demands of unaccommodated users need not be enforced.

The formulation in (19) is non-convex because of the rank-1 constraint (19h), even though the objective in (19a) and the constraints in (19b)–(19g) are convex in $\mathbf{\Psi}$ and $\{t_u\}$. Hence, apart from the rank-1 constraint, the formulation in (19) can be efficiently solved over PSD matrices, and it is this observation that forms the basis of the proposed SDR-based technique.

The philosophy of the SDR-based technique is to convert the deterministic optimization problem in (19) into a stochastic one by treating the PSD matrix $\mathbf{\Psi}$ generated by (19) as the covariance matrix of a multivariate Gaussian distribution in the prospective stochastic formulation [27]. The technique proceeds by drawing random samples from the aforementioned distribution. The samples are rounded to obtain candidate antipodal solutions and the feasible candidate that yields the largest objective is selected by the algorithm.

We conclude our presentation of the proposed SDR-based technique with a discussion on the choice of σ in (19f) and its impact on the performance. To begin with, we note that in this technique, candidate solutions are obtained by quantizing samples drawn from the multivariate Gaussian distribution. Now, if σ is large, users with normalized aggregate rate close to zero will be treated as if they were accommodated in the relaxed formulation in the (11c) and (11d) constraints in (16b). However, in the subsequent randomization step, such users are likely to be not accommodated. Hence, if σ is chosen to be too large, it would result in an overly restrictive formulation wherein users that are not accommodated in the network may enforce their QoS constraints in the relaxed formulation. Hence, it can be seen that the value of σ must be carefully chosen in order to ensure that, when the Gaussian sample vectors generated via the solution of (16), are quantized, the constraints are only enforced for those users who are actually accommodated in the network. Finding an appropriate value for σ analytically appears to be difficult. Hence, such a value will be obtained numerically in Section VI.

1) Computational Complexity Analysis: We first examine the complexity required for solving the relaxed version of (19), i.e., when the rank-1 constraint is dropped. The relaxation of (19) is solved only once, and the complexity of solving it, which is a PSD-constrained convex problem, is obtained using an approach similar to the one given in Section V-A.1, and requires $\mathcal{O}\left((BUSL + 2U)^{3.5}\right)$. For the Gaussian randomization procedure that is used after solving the relaxation of (19), the complexity of generating and evaluating the objective function corresponding to the J random samples is $\mathcal{O}\left((BUSL)^2 J\right)$ [24]. Combining these observations, the complexity of the algorithm is $\mathcal{O}\left((BUSL + 2U)^{3.5} + (BUSL)^2 J\right)$. Note that the number of randomization samples does not affect the *grosso modo* complexity order [28]. This complexity is significantly less than that of exhaustive search one. The latter is $\mathcal{O}(2^{BUSL})$.

An important observation is that the SDR-based technique relies on constructing PSD matrices with a number of entries that scales with the square of the number of variables. For the case without RB-reuse, this number is $BUSL$ and for the case with RB-reuse of $q = 1$ this number is $BUSL(1 + (B-1)L)$. Hence, it can be seen that, despite being polynomial, the complexity of the SDR-based technique grows fast with the size of the network, rendering simulation of large networks a challenging task, especially for the case with RB-reuse. In other words, while our algorithm is fundamentally more efficient to implement than exhaustive search, implementing it for large networks can be computationally prohibitive, at least on non-special-purpose computers. Hence, the formulation proposed in our work is readily applicable in small-to-medium networks. For larger networks, the given formulation can be used to develop decentralized designs, which are likely to be more practical in those cases.

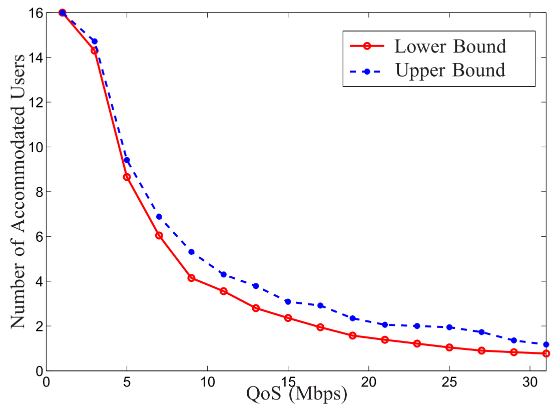
VI. Numerical Results

In this section, we investigate the performance of the proposed joint design approaches, and provide numerical examples to illustrate the merits of these approaches for network instances with and without time-sharing and with and without RB-reuse.[3] We consider the 3GPP propagation model in [30] in a two-tier HetNet scenario with one macro BS with a maximum transmit power of 46 dBm, and several pico BSs with a maximum transmit power of 35 dBm. The location of the macro BS is assumed to be fixed, whereas the locations of the pico BSs and the users are assumed to be randomly distributed. The users have identical QoS demands, and the power levels used by the b-th BS are taken to be L equally spaced points in the interval $[0.05P_b^{\max},\ 0.5P_b^{\max}]$, unless otherwise stated. As per the 3GPP propagation model in [30], shadowing is assumed to have a log-normal distribution with a standard deviation of 8 dB for the macro BS, and 10 dB for the pico BSs. Denoting the distance between users and BSs by d (km), we express the path losses for the link between the macro BS and the users, and the link between the pico BSs and the users by $PL(d) = 128.1 + 37.6\log_{10}(d)$, and $PL(d) = 140.7 + 36.7\log_{10}(d)$, respectively. The bandwidth of each RB is assumed to be 180 kHz, and the noise power spectral density at all receivers is assumed to be -174 dBm/Hz.

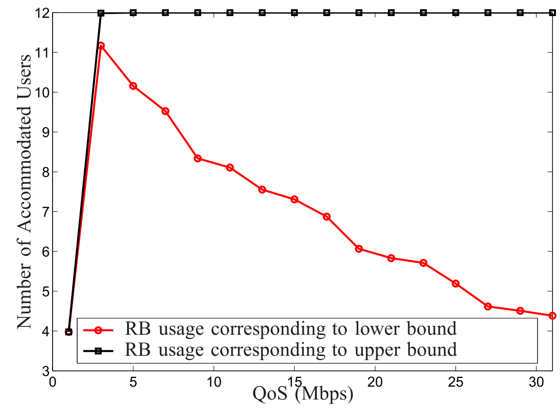
For all simulations, results are averaged over 100 independent channel realizations, and the mathematical programs are solved using the CVX package [31] with the SeDuMi solver. The number of Gaussian samples used in the randomization phase of the SDR programs is set to be $J = 10^4$. For the case with time-sharing, the value of σ in (13) is chosen to be 100, whereas for the case without time-sharing, numerical investigations suggested that setting $\sigma \in [1.5, 2]$ in (16) tends to generate favourable results; in our simulations we set $\sigma = 1.8$.

Example 1 (Tightness of Bounds—Time-Sharing Case): In this example, we consider a time-sharing scenario without

[3] Preliminary results on this problem were reported in [29] for a simplified variant of the joint design problem with neither time-sharing nor RB-reuse and with fixed RBs and power allocations.



(a) Upper and lower bounds on the average number of users.



(b) Average aggregate RB usage.

Fig. 1. Upper and lower bounds for the network with time-sharing and without frequency-reuse.

TABLE I
RATES IN RB-REUSE CASE

	(u_1, ℓ_1, n_1)	(u_1, ℓ_1, n_2)	(u_1, ℓ_2, n_1)	(u_1, ℓ_2, n_2)	(u_2, ℓ_1, n_1)	(u_2, ℓ_1, n_2)	(u_2, ℓ_2, n_1)	(u_2, ℓ_2, n_2)
(b_1, s_1, k_2)	4.6390	4.6280	5.0570	5.0459	0.0001	0.0001	0.0003	0.0001
(b_1, s_2, k_2)	5.1236	5.1235	5.5415	5.5415	0.0001	0.0000	0.0002	0.0001
(b_2, s_1, k_1)	0.0001	0.0000	0.0003	0.0000	2.6915	2.2788	3.1094	2.6967
(b_2, s_2, k_1)	0.0002	0.0000	0.0005	0.0000	2.0166	1.6037	2.4344	2.0212

RB-reuse in a network with $B = 5$ BSs, one macro and four picos, $U = 16$ users, $S = 12$ RBs, and $L = 3$ power levels, yielding a total of 2880 variables.

Fig. 1(a) investigates the tightness of the bounds generated by the algorithms in Section V-A. From this figure, it can be seen that the gap between the bounds is generally small, especially at low QoS values. For instance, for QoS values below 1 Mbps, the bounds coincide. For higher QoS values, the maximum gap between the bounds is about 0.9 user on average.

To complement the preceding discussion, in Fig. 1(b), we investigate the aggregate usage of the RBs corresponding to the aforementioned bounds. From this figure, it can be seen that, although the bounds on the number of users are tight, the difference in the corresponding RB usage can be relatively large, especially at high QoS demands. To understand this phenomenon, we note that the cost of accommodating an extra user increases with the QoS demand. Hence, at high QoS demands, the difference between the number of users generated by the upper and lower bounds corresponds to a large gap between the number of RBs required to realize these demands. This gap becomes smaller when the gap between the bounds on the number of users is small and disappears when the bounds coincide. For instance, for QoS demands less than 1 Mbps, the bounds on the number of users in Fig. 1(a) coincide and so do the corresponding RB usages in Fig. 1(b). In other words, the gap between the bounds on RB usage scales with QoS demands and the gap between the bounds on accommodated users. □

Example 2 (Opportunistic RB-Reuse): In this example, we provide and discuss the solution generated by the approach proposed in Section V-A for a particular realization of a network with time-sharing, 2 BSs (b_1 and b_2), 2 users (u_1, and u_2), 2 RBs (s_1 and s_2), and 2 power levels (ℓ_1 and ℓ_2). We construct two tables for the data rate that can be communicated by each user-to-BS association on each RB: Table I pertains to the case with RB-reuse. In this table, k_i refers to the BS interfering with the transmission of BS b_j, and n_i refers to the power level used by BS k_i, $i, j \in \{1, 2\}$, $i \neq j$. Table II pertains to the case with no RB-reuse.

TABLE II
RATES IN NO RB-REUSE CASE

	(u_1, ℓ_1)	(u_1, ℓ_2)	(u_2, ℓ_1)	(u_2, ℓ_2)
(b_1, s_1)	4.6418	5.0598	0.9595	1.3722
(b_1, s_2)	5.1236	5.5416	0.9891	1.4024
(b_2, s_1)	0.0028	0.0139	3.6510	4.0689
(b_2, s_2)	0.0000	0.0001	3.0056	3.4235

The rate values in these tables are in Mbps and obtained using the expression in (2) with b_1 and b_2 being macro and pico BSs, respectively. If the users' QoS demands are set to 6 Mbps, one can verify that two users cannot be accommodated when both RBs are reused or neither of them is reused over the entire signalling interval. In particular, the fact that no entries in Table II exceeds 6 Mbps directly implies that aggregate RB usage of each user should be greater than 1 to be accommodated. However, since the aggregate RB usage in the network cannot be greater than 2, the two users cannot be accommodated in the no RB-reuse case. In a similar fashion, Table I shows that the sum of any two entries corresponding to the u_2-th user cannot add up to 6 Mbps. Hence, user u_2 cannot be accommodated if RB-reuse is enforced on all RBs. Next, we will show that in this example, both users can be accommodated if the

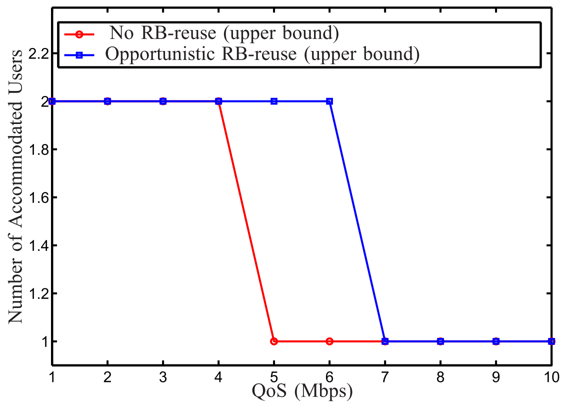


Fig. 2. Opportunistic RB-reuse on the number of users.

RBs are allowed to be reused opportunistically. For instance, these users can be accommodated if the first RB, s_1, is reused for 25% of time, and not reused for 75% of time, and the second RB, s_2, is reused over the entire signalling interval. In this case, $y_{b_1u_1}^{s_1\ell_1k_2n_2} = 0.25$, $y_{b_1u_1}^{s_2\ell_1k_2n_2} = 1$, $y_{b_2u_2}^{s_1\ell_2k_1n_1} = 0.25$, $y_{b_2u_2}^{s_2\ell_2k_1n_1} = 1$ and $y_{b_2u_2}^{s_1\ell_2} = 0.75$. Using these values yields a QoS of 6.2805 Mbps for user u_1 and 6.2634 Mbps for user u_2. The corresponding aggregate RB usage is 3.25.

To confirm the above observations, we perform the optimization proposed in Section V-A. This optimization yielded the following solution: In the RB-reuse case, we have $y_{b_1u_1}^{s_1\ell_1k_2n_2} = 0.1770$, $y_{b_1u_1}^{s_2\ell_1k_2n_2} = 0.8630$, $y_{b_2u_2}^{s_1\ell_2k_1n_1} = 0.1770$, and $y_{b_2u_2}^{s_2\ell_2k_1n_1} = 0.8630$, and in the no RB-reuse case, we have $y_{b_1u_1}^{s_2\ell_2} = 0.1370$ and $y_{b_2u_2}^{s_1\ell_2} = 0.8230$. Since the proposed algorithm opportunistically switches between the RB-reuse case and the no RB-reuse one, this algorithm offers valuable design flexibility, and enables the two users to be accommodated in the network. The QoS of the users generated by the algorithm in this case is 6 Mbps per users and the aggregate RB usage is 3.04. Hence, it can be seen that the algorithm ensures that the users' QoS demands while minimizing aggregate RB usage. Fig. 2 provides a comparison between the algorithms with opportunistic RB-reuse and without RB-reuse for the rate values given in Table I and Table II for a range of QoS values. From this figure, it can be seen that the algorithm with opportunistic RB-reuse shows a superior performance for QoS values between 5 and 6 Mbps. □

Example 3 (Tightness of Bounds—No Time-Sharing Case): In this example, we compare the lower bound on the number of users generated by the algorithm in Section V-B with the optimal number of accommodated users obtained by exhaustive search for the case with no time-sharing and no RB-reuse. Since exhaustive search is computationally expensive for large networks, we consider a relatively small network with $B = 2$ BSs, one macro and one pico, $U = 3$ users, $S = 4$ RBs, and $L = 1$ power level of $\{0.25P_b^{\max}\}$. The optimal number of users obtained by exhaustive search and the lower bound generated by the algorithm in Section V-B are depicted in Fig. 3(a). From this figure, it can be seen that the technique proposed in Section V-B yields a close-to-optimal user-to-BS association and RBs allocations especially at low QoS values. For instance, for QoS demands below 3 Mbps, the average number of users accommodated by the proposed technique coincides with that accommodated by exhaustive search. For QoS demands between 3 and 7 Mbps, the average number of users accommodated by the proposed technique is about 75% of that accommodated by exhaustive search. It is worth noting that the gap between the maximum number of accommodated users and the one generated by the SDR-based technique is at most 1 user. In fact, for low QoS the gap is zero and for high QoS numerical experiments indicated that this gap does not exceed 1 user, i.e., the smallest nonzero gap. Although we have not been able to develop an analytical proof of the latter statement, numerical analysis confirms that the proposed SDR-based technique can yield close-to-optimal performance.

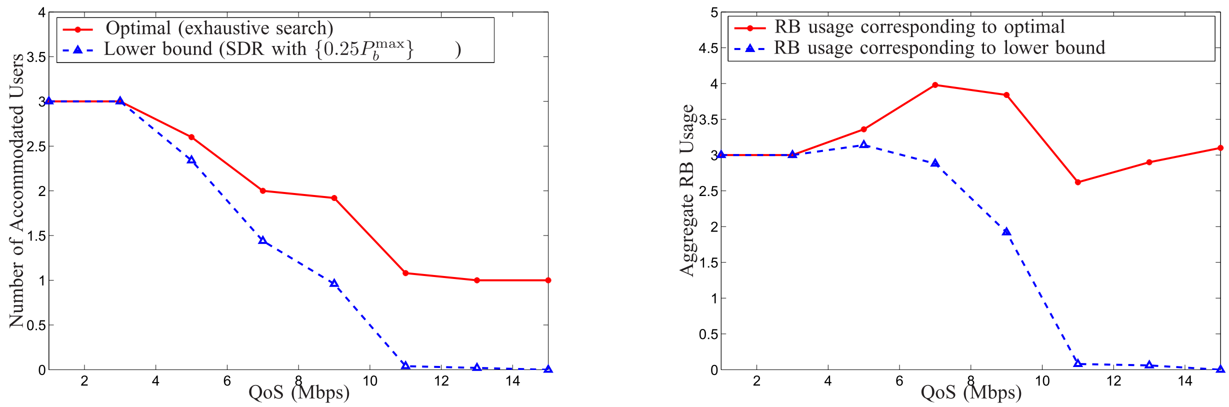
Analogous to the case in Example 1, we now investigate the aggregate usage of the RBs in the network without time-sharing. In Fig. 3(b), we compare the RB usage when both exhaustive search and the algorithm proposed in Section V-B are used. Similar to the case in Example 1, we note that, as the difference between the optimal number of users and the lower bound in Fig. 3(a) increases, the gap between the corresponding RB usages in Fig. 3(b) widens; a zero difference in the number of accommodated users corresponds to a zero RB usage gap. However, unlike Fig. 1(b), Fig. 3(b) exhibits a non-monotonic behaviour. This is because in this example time-sharing is not allowed. This implies that the association and allocation variables are discrete, which bars the users from occupying fractions of the signalling interval. Hence, for some QoS demands, there maybe extra RBs that, although not used, do not suffice to accommodate additional users. As the QoS demand increases, these extra RBs can be used by the existing users, which explains the rise in the RB usage curves observed at a QoS demand of 5 Mbps. As the QoS continues to increase, less users can be accommodated and the RB drops to zero.

As in Example 2, we now compare the optimal solution with the solution generated by the SDR-based technique for a network realization with 2 BSs (b_1 and b_2), 3 users (u_1, u_2 and u_3) and 4 RBs ($s_1, \ldots, s_4$). We use (2) to construct the table of the data rates (in Mbps) that can be communicated for every user-to-BS association on each RB, cf. the first table in (20), as shown at the bottom of the next page.

If the users' QoS demands are set to 3 Mbps, one can readily verify that the 3 users can be accommodated with user u_1 associated with BS b_1 on any one of the four RBs, i.e., s_1, s_2, s_3, and s_4, user u_2 can be associated with BS b_2 on any one of the remaining three RBs, and user u_3 can be associated with BS b_1 on any pair of the remaining two RBs.

A confirmation of the above observations is obtained by performing exhaustive search, which yielded the association table in the middle of (20). The corresponding matrix generated by the proposed SDR-based technique is given in the association table on the right of (20). Comparing the association tables, it can be seen that in both tables, the users are associated with the same BSs using the same number of RBs, albeit with different ones. In both tables, the number of RBs exhausted to accommodate the users is 4. □

Example 4 (Impact of BS Transmit Powers—No Time-Sharing Case): In this example, we investigate the performance of the SDR-based technique in a network with



(a) Optimal and lower bound on the average number of users. (b) Corresponding average RB usage

Fig. 3. Optimal and lower bound for the network with no time-sharing and no RB-reuse.

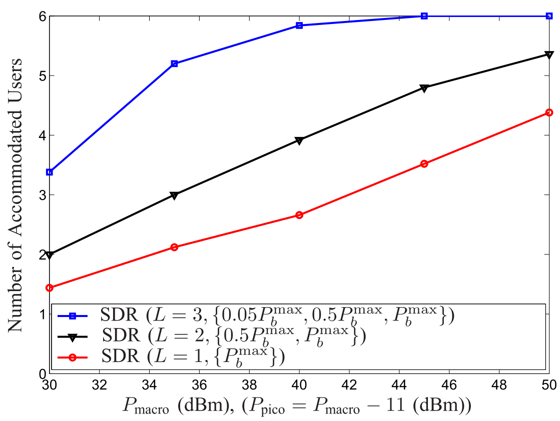


Fig. 4. Impact of the BS transmit power on performance.

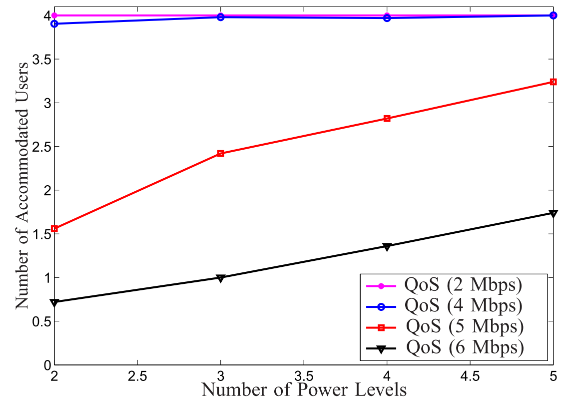


Fig. 5. Impact of the number of power levels on performance.

no time-sharing, no RB-reuse and with $B = 5$ BSs, one macro and two picos, $U = 6$ users, $S = 10$ RBs and $Q_u = 2$ Mbps, $\forall u$. For this network we plot in Fig. 4 the number of accommodated users versus the transmit power of the macro BS, P_{macro}, assuming that $P_{\text{pico}} = P_{\text{macro}} - 11$ dBm.

From Fig. 4, it can be seen that increasing the number of power levels improves the network performance. For instance, for all transmit powers, the algorithm with $L = 3$ power levels shows a performance advantage of 15 dBm over the case with $L = 2$ power levels for an accommodated number of users of $U = 5$. □

Example 5 (Impact of Power Levels—No Time-Sharing Case): In this example, we investigate the impact of the number of power levels on the performance of the SDR-based technique. To do so, we consider a no time-sharing scenario with no RB-reuse in a network with $B = 2$ BSs, one macro and one pico, $U = 4$ users and $S = 6$ RBs. In Fig. 5, we plot the number of accommodated users versus the number of power levels for various QoS demands.

From this figure, it can be seen that for low QoS demands, there is no perceptible change in the number of accommodated users with the increase of the number of power level. This is because, for low QoS demands, there are enough RBs to support all users. However, as the QoS demands increase, the increase in the number of power levels yields a large increase in the number of users accommodated in the network. For instance, for a QoS of 5 Mbps, the number of accommodated users doubles as the number of power levels increases from two levels to four. Our simulations suggest that this phenomenon continues to prevail for larger networks. □

	u_1	u_2	u_3
(b_1, s_1)	3.0463	2.0456	2.2549
(b_1, s_2)	3.7359	1.5271	2.8976
(b_1, s_3)	3.0241	1.9794	2.2589
(b_1, s_4)	3.3721	1.2696	2.7857
(b_2, s_1)	0.1003	3.8978	0.0003
(b_2, s_2)	0.6106	4.7605	0.0001
(b_2, s_3)	0.0505	4.6056	0.0002
(b_2, s_4)	0.1971	4.4528	0.0001

	u_1	u_2	u_3
(b_1, s_1)	1	0	0
(b_1, s_2)	0	0	1
(b_1, s_3)	0	0	1
(b_1, s_4)	0	0	0
(b_2, s_1)	0	0	0
(b_2, s_2)	0	0	0
(b_2, s_3)	0	0	0
(b_2, s_4)	0	1	0

	u_1	u_2	u_3
(b_1, s_1)	1	0	0
(b_1, s_2)	0	0	1
(b_1, s_3)	0	0	0
(b_1, s_4)	0	0	1
(b_2, s_1)	0	0	0
(b_2, s_2)	0	0	0
(b_2, s_3)	0	1	0
(b_2, s_4)	0	0	0

(20)

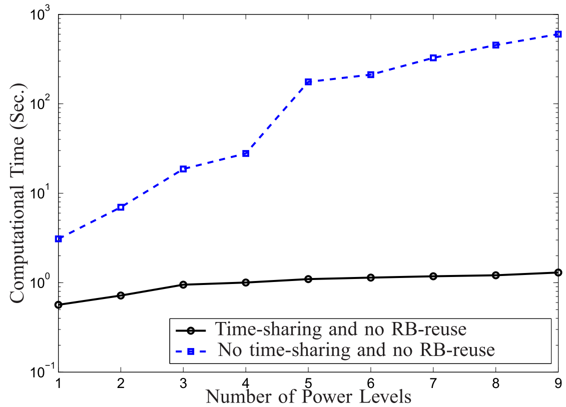


Fig. 6. Computational time versus network size.

Example 6 (Computational Time): We consider a no RB-reuse scenario with and without time-sharing in a network with $B = 3$ BSs, one macro and two picos, $U = 4$ users, $S = 6$ RBs, and $Q_u = 2$ Mbps, $\forall u$. In this setting, we vary the number of power levels from 1 to 9.

Fig. 6 shows the average computational time required by the algorithms proposed in Section V-A and Section V-B against the number of power levels. From this figure, it can be seen that both algorithms require a polynomial amount of computational time, and the algorithm proposed in Section V-A requires much lower computational time than the one proposed in Section V-B. $\square$

VII. Conclusion

In this paper, we considered the joint optimization of the user-to-BS associations, and the RBs and the power allocations in HetNets. We investigated this optimization for two novel network instances with opportunistic RB-reuse: one in which the RBs can be reused and the user-to-BS associations and the power allocations can be time-shared throughout the signalling interval, and one in which the RBs can be reused but the user-to-BS associations and the power allocations cannot be time-shared. Unfortunately, both design problems are non-convex, and therefore, difficult to solve. To deal with this difficulty, we provided a formulation, in which the discrete number of users is approximated by a continuous function. We then developed tight upper and lower bounds for the time-shared version of this formulation, and for the no time-shared one we developed a lower bound using the SDR with randomization technique. Numerical results confirm that the developed bounds yield close-to-optimal performance.

Appendix A
Proof of Proposition 1

The proof of this proposition uses a contradiction argument analogous to the one in [32] and [17] but for a different design objective, network functionalities, and network instances.

Let $\{\tilde{y}_{bu}^{s\ell kn}\}$ be the optimal solution for the objective in Section III-B.3 and $\forall b, u, s, \ell, k, n$, let $\{\hat{y}_{bu}^{s\ell kn}\}$ be a feasible solution which accommodates one user more than the optimal solution, i.e., $f(\hat{\mathbf{y}}) = f(\tilde{\mathbf{y}}) + 1$, where

$$f(\hat{\mathbf{y}}) = \sum_{u\in\mathcal{U}}\Big(1 - e^{-\frac{\sigma}{Q_u}\sum_{b\in\mathcal{B}}\sum_{s\in\mathcal{S}}\sum_{\ell\in\mathcal{P}_b}\hat{y}_{bu}^{s\ell 00}R_{bu}^{s\ell 00}+\sum_{k\in\mathcal{B}\backslash b}\sum_{n\in\mathcal{P}_k}\hat{y}_{bu}^{s\ell kn}R_{bu}^{s\ell kn}}\Big),$$

and

$$f(\tilde{\mathbf{y}}) = \sum_{u\in\mathcal{U}}\Big(1 - e^{-\frac{\sigma}{Q_u}\sum_{b\in\mathcal{B}}\sum_{s\in\mathcal{S}}\sum_{\ell\in\mathcal{P}_b}\tilde{y}_{bu}^{s\ell 00}R_{bu}^{s\ell 00}+\sum_{k\in\mathcal{B}\backslash b}\sum_{n\in\mathcal{P}_k}\tilde{y}_{bu}^{s\ell kn}R_{bu}^{s\ell kn}}\Big).$$

Then, the objective function of the feasible solution can be expressed as

$$\begin{aligned}\rho f(\hat{\mathbf{y}}) - (1-\rho)g(\hat{\mathbf{y}}) &\geq \rho f(\tilde{\mathbf{y}}) + \rho - 2(1-\rho)S\\ &\geq \rho f(\tilde{\mathbf{y}})\\ &\geq \rho f(\tilde{\mathbf{y}}) - (1-\rho)g(\tilde{\mathbf{y}}),\end{aligned} \tag{21}$$

where the aggregate usage of RBs in the network for the feasible and optimal solutions are respectively given by $g(\hat{\mathbf{y}}) = \sum_{u\in\mathcal{U}}\sum_{b\in\mathcal{B}}\sum_{s\in\mathcal{S}}\sum_{\ell\in\mathcal{P}_b}\big(\sum_{k\in\mathcal{B}\backslash b}\sum_{n\in\mathcal{P}_k}\hat{y}_{bu}^{s\ell kn} + \hat{y}_{bu}^{s\ell 00}\big)$, and $g(\tilde{\mathbf{y}}) = \sum_{u\in\mathcal{U}}\sum_{b\in\mathcal{B}}\sum_{s\in\mathcal{S}}\sum_{\ell\in\mathcal{P}_b}\big(\tilde{y}_{bu}^{s\ell 00} + \sum_{k\in\mathcal{B}\backslash b}\sum_{n\in\mathcal{P}_k}\tilde{y}_{bu}^{s\ell kn}\big)$. Now, the total usage of RBs in the system can be at most $2S$. Supposing that $\rho - 2(1-\rho)S > 0$, i.e., $\rho > \frac{2S}{1+2S}$, it can be seen that the value of the objective function of the feasible solution is greater than the value of the objective at the optimal solution, that is, $\rho f(\hat{\mathbf{y}}) - (1-\rho)g(\hat{\mathbf{y}}) \geq \rho f(\tilde{\mathbf{y}}) - (1-\rho)g(\tilde{\mathbf{y}})$. This contradicts the optimality of $\{\tilde{y}_{bu}^{s\ell kn}\}$, and hence, when $\rho \in \big(\frac{2S}{1+2S}, 1\big)$, the composite objective ensures that the maximum number of users is accommodated with the least number of RBs.

Appendix B
Proof of Proposition 2

A function $f : \mathbb{R}^n \to \mathbb{R}$ is self-concordant if it satisfies $\left|\frac{\partial^3 f(x+kv)}{\partial k^3}\right| \leq 2\frac{\partial^2}{\partial k^2}f(x+kv)^{3/2}$, for all $x \in \mathbf{dom}\ f$ and for all v [33]. Now, the log-barrier function for (13) is given by

$$\Omega = k\Big(\rho\sum_{u\in\mathcal{U}}(1-t_u) - (1-\rho) \times \sum_{s\in\mathcal{S}}\sum_{b\in\mathcal{B}}\sum_{u\in\mathcal{U}}\sum_{\ell\in\mathcal{P}_b}\Big(y_{bu}^{s\ell 00} + \sum_{k\in\mathcal{B}\backslash b}\sum_{n\in\mathcal{P}_k}y_{bu}^{s\ell kn}\Big)\Big) + \phi, \tag{22}$$

where ϕ is the logarithmic barrier defined as

$$\begin{aligned}\phi = &-\sum_{b\in\mathcal{B}}\log\Big(P_b^{\max} - \sum_{u\in\mathcal{U}}\sum_{s\in\mathcal{S}}\sum_{\ell\in\mathcal{P}_b}\Big(y_{bu}^{s\ell 00} + \sum_{k\in\mathcal{B}\backslash b}\sum_{n\in\mathcal{P}_k}y_{bu}^{s\ell kn}\Big)p_b^{\ell}\Big)\\ &-\sum_{u\in\mathcal{U}}\log\Big(\sum_{b\in\mathcal{B}}\sum_{s\in\mathcal{S}}\sum_{\ell\in\mathcal{P}_b}\Big(y_{bu}^{s\ell 00}R_{bu}^{s\ell 00} + \sum_{k\in\mathcal{B}\backslash b}\sum_{n\in\mathcal{P}_k}y_{bu}^{s\ell kn}R_{bu}^{s\ell kn}\Big)\\ &- Q_u(1-t_u)\Big) - \sum_{s\in\mathcal{S}}\log\Big(1 - \sum_{b\in\mathcal{B}}\sum_{u\in\mathcal{U}}\sum_{\ell\in\mathcal{P}_b}y_{bu}^{s\ell 00}\\ &- \frac{1}{2}\sum_{b\in\mathcal{B}}\sum_{u\in\mathcal{U}}\sum_{\ell\in\mathcal{P}_k}\sum_{k\in\mathcal{B}\backslash b}\sum_{n\in\mathcal{P}}y_{bu}^{s\ell kn}\Big)\end{aligned}$$

$$- \sum_{b\in\mathcal{B}}\sum_{u\in\mathcal{U}}\sum_{s\in\mathcal{S}}\sum_{\ell\in\mathcal{P}_b}\Big(\sum_{k\in\mathcal{B}\setminus b}\sum_{n\in\mathcal{P}_k}\log\Big(\sum_{\tilde{u}\in\mathcal{U}\setminus u} y_{k\tilde{u}}^{snb\ell} - y_{bu}^{s\ell kn}\Big)$$
$$+ \log\Big(1 - y_{bu}^{s\ell 00} + \sum_{k\in\mathcal{B}\setminus b}\sum_{n\in\mathcal{P}_k} y_{bu}^{s\ell kn}\Big)\Big)$$
$$- \sum_{u\in\mathcal{U}}\sum_{s\in\mathcal{S}}\log\Big(1 - \sum_{b\in\mathcal{B}}\sum_{\ell\in\mathcal{P}_b}\Big(y_{bu}^{s\ell 00} - \sum_{k\in\mathcal{B}\setminus b}\sum_{n\in\mathcal{P}_k} y_{bu}^{s\ell kn}\Big)\Big)$$
$$- \sum_{b\in\mathcal{B}}\sum_{s\in\mathcal{S}}\log\Big(1 - \sum_{\ell\in\mathcal{P}_b}\sum_{u\in\mathcal{U}}\Big(y_{bu}^{s\ell 00} - \sum_{k\in\mathcal{B}\setminus b}\sum_{n\in\mathcal{P}_k} y_{bu}^{s\ell kn}\Big)\Big)$$
$$- \sum_{b\in\mathcal{B}}\sum_{u\in\mathcal{U}}\sum_{s\in\mathcal{S}}\sum_{\ell\in\mathcal{P}_b}\sum_{k\in\{0\}\cup\mathcal{B}\setminus b}\sum_{n\in\mathcal{P}_k}\log\Big(y_{bu}^{s\ell kn}\Big)$$
$$- \sum_{u\in\mathcal{U}}\log\Big(\log(t_u) + \frac{\sigma}{Q_u}\sum_{b\in\mathcal{B}}\sum_{s\in\mathcal{S}}\sum_{\ell\in\mathcal{P}_b}\big(y_{bu}^{s\ell 00}R_{bu}^{s\ell 00}$$
$$+ \sum_{k\in\mathcal{B}\setminus b}\sum_{n\in\mathcal{P}_k} y_{bu}^{s\ell kn}R_{bu}^{s\ell kn}\big)\Big) - \log(t_u). \tag{23}$$

Self-concordance of the first eight terms can be readily verified. For the remaining two terms, we utilize a known self-concordant form, $-\log(\log(a)+b)-\log(a)$ [33], where $a = \log(t_u)$, and $b = \frac{\sigma}{Q_u}\sum_{b\in\mathcal{B}}\sum_{s\in\mathcal{S}}\sum_{\ell\in\mathcal{P}_b}\big(y_{bu}^{s\ell 00}R_{bu}^{s\ell 00} + \sum_{k\in\mathcal{B}\setminus b}\sum_{n\in\mathcal{P}_k} y_{bu}^{s\ell kn}R_{bu}^{s\ell kn}\big)$, which shows that Ω is self-concordant. Using self-concordance of the log-barrier function, the number of Newton steps required to solve (13) can be shown to be proportional to $\sqrt{n}$, where n is the number of inequality constraints. Since each Newton step requires cubic complexity, the complexity of solving (13) is proportional to $n^{3.5}$, which establishes the claim of the proposition.

Acknowledgements

The authors would like to acknowledge the insightful and constructive feedback they received from the Associate Editor and the anonymous Reviewers.

References

[1] H. Boostanimehr and V. K. Bhargava, "Unified and distributed QoS-driven cell association algorithms in heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 3, pp. 1650–1662, Mar. 2015.

[2] 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.

[3] G. Hegde, O. D. Ramos-Cantor, Y. Cheng, and M. Pesavento, "Optimal resource block allocation and muting in heterogeneous networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2016, pp. 3581–3585.

[4] S. Sesia, I. Toufik, and M. Baker, *LTE—The UMTS Long Term Evolution*. Hoboken, NJ, USA: Wiley, 2009.

[5] H. Zhang, L. Venturino, N. Prasad, P. Li, S. Rangarajan, and X. Wang, "Weighted sum-rate maximization in multi-cell networks via coordinated scheduling and discrete power control," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 6, pp. 1214–1224, Jun. 2011.

[6] R. Rashtchi, R. H. Gohary, and H. Yanikomeroglu, "Generalized cross-layer designs for generic half-duplex multicarrier wireless networks with frequency-reuse," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 458–471, Jan. 2016.

[7] 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.

[8] D. Bethanabhotla, O. Y. Bursalioglu, H. C. Papadopoulos, and G. Caire, "Optimal user-cell association for massive MIMO wireless networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 1835–1850, May 2016.

[9] F. Boccardi *et al.*, "Why to decouple the uplink and downlink in cellular networks and how to do it," *IEEE Commun. Mag.*, vol. 54, no. 3, pp. 110–117, Mar. 2016.

[10] A. T. Gamage, H. Liang, and X. Shen, "Two time-scale cross-layer scheduling for cellular/WLAN interworking," *IEEE Trans. Commun.*, vol. 62, no. 8, pp. 2773–2789, Aug. 2014.

[11] K. Son, S. Chong, and G. Veciana, "Dynamic association for load balancing and interference avoidance in multi-cell networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 7, pp. 3566–3576, Jul. 2009.

[12] I. Guvenc, M.-R. Jeong, I. Demirdogen, B. Kecicioglu, and F. Watanabe, "Range expansion and inter-cell interference coordination (ICIC) for picocell networks," in *Proc. IEEE Veh. Tech. Conf. (VTC-Fall)*, Sep. 2011, pp. 1–6.

[13] K. Shen and W. Yu, "Distributed pricing-based user association for downlink heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1100–1113, Jun. 2014.

[14] E. Aryafar, A. Keshavarz-Haddad, M. Wang, and M. Chiang, "RAT selection games in HetNets," in *Proc. IEEE INFOCOM*, Apr. 2013, pp. 998–1006.

[15] M. Hong and A. Garcia, "Mechanism design for base station association and resource allocation in downlink OFDMA network," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 11, pp. 2238–2250, Dec. 2012.

[16] S. Corroy and R. Mathar, "Semidefinite relaxation and randomization for dynamic cell association in heterogeneous networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2012, pp. 2373–2378.

[17] L. P. Qian, Y. J. A. Zhang, Y. Wu, and J. Chen, "Joint base station association and power control via benders' decomposition," *IEEE Trans. Wireless Commun.*, vol. 12, no. 4, pp. 1651–1665, Apr. 2013.

[18] R. Sun, M. Hong, and Z.-Q. Luo, "Joint downlink base station association and power control for max-min fairness: Computation and complexity," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 6, pp. 1040–1054, Jun. 2015.

[19] V. N. Ha and L. B. Le, "Distributed base station association and power control for heterogeneous cellular networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 1, pp. 282–296, Jan. 2014.

[20] R. Madan, J. Borran, A. Sampath, N. Bhushan, A. Khandekar, and T. Ji, "Cell association and interference coordination in heterogeneous LTE-A cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1479–1489, Dec. 2010.

[21] Q. Kuang, J. Speidel, and H. Droste, "Joint base-station association, channel assignment, beamforming and power control in heterogeneous networks," in *Proc. IEEE Veh. Tech. Conf. (VTC-Spring)*, May 2012, pp. 1–5.

[22] C. S. Chen, F. Baccelli, and L. Roullet, "Joint optimization of radio resources in small and macro cell networks," in *Proc. IEEE VTC Spring*, Budapest, Hungary, May 2011, pp. 1–5.

[23] G. Li and H. Liu, "Resource allocation for OFDMA relay networks with fairness constraints," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 11, pp. 2061–2069, Nov. 2006.

[24] Z.-Q. Luo, W.-K. Ma, A. M.-C. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 20–34, May 2010.

[25] J. F. Sturm, "Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones," *Optim. Methods Softw.*, vol. 11, nos. 1–4, pp. 625–653, 1999.

[26] M. X. Goemans and D. P. Williamson, "Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming," *J. Assoc. Comput. Mach.*, vol. 42, no. 6, pp. 1115–1145, 1995.

[27] A. Wiesel, Y. C. Eldar, and S. Shamai (Shitz), "Semidefinite relaxation for detection of 16-QAM signaling in MIMO channels," *IEEE Signal Process. Lett.*, vol. 12, no. 9, pp. 653–656, Sep. 2005.

[28] N. D. Sidiropoulos and Z. Q. Luo, "A semidefinite relaxation approach to MIMO detection for high-order QAM constellations," *IEEE Signal Process. Lett.*, vol. 13, no. 9, pp. 525–528, Sep. 2006.

[29] H. U. Sokun, R. H. Gohary, and H. Yanikomeroglu, "QoS-guaranteed user association in HetNets via semidefinite relaxation," in *Proc. IEEE 82nd Veh. Technol. Conf. (VTC-Fall)*, Sep. 2015, pp. 1–5.

[30] *Evolve Universal Terrestrial Radio Access (E-UTRA); Mobility Enhancements in Heterogeneous Networks v11.1.0*, document TR 36.839, 3GPP, Jan. 2013.

[31] M. Grant and S. Boyd. (Mar. 2014). *CVX: MATLAB Software for Disciplined Convex Programming, Version 2.1*. [Online]. Available: http://cvxr.com/cvx

[32] E. Matskani, N. D. Sidiropoulos, Z. Q. Luo, and L. Tassiulas, "Convex approximation techniques for joint multiuser downlink beamforming and admission control," *IEEE Trans. Wireless Commun.*, vol. 7, no. 7, pp. 2682–2693, Jul. 2008.

[33] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

Hamza Umit Sokun (S'14) received the B.Sc. degree (Hons.) in electronics engineering from Kadir Has University, Turkey, in 2010, and the M.Sc. degree in electrical and electronics engineering from Ozyegin University, Turkey, in 2012. He is currently pursuing the Ph.D. degree in electrical and computer engineering with Carleton University, Canada. His research interests include cooperative communications, green communications, radio resource allocation, and applications of optimization for wireless communications.

Ramy H. Gohary (S'02–M'06–SM'13) received the B.Eng. degree (Hons.) from Assiut University, Egypt, in 1996, the M.Sc. degree from Cairo University, Egypt, in 2000, and the Ph.D. degree from McMaster University, Hamilton, ON, Canada, in 2006, all in electronics and communications engineering. He is currently an Assistant Professor with the Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada. His research interests include machine-to-machine communications, Internet of Things, cross-layer design of wireless networks, analysis and design of MIMO wireless communication systems, applications of optimization and geometry in signal processing and communications, information theoretic aspects of multiuser communication systems, and applications of iterative detection and decoding techniques in multiple antenna and multiuser systems. He is also interested in the analysis, design, and hardware implementation of computer networks. He is a member of the College of Reviewers and the Panel Review Board of the Ontario Centre of Excellence. He was named the Best Professor by the Carleton Student Engineering Society in 2015. He is a registered Limited Engineering Licensee in the province of Ontario, Canada.

Halim Yanikomeroglu (F'07) was born in Giresun, Turkey, in 1968. He received the B.Sc. degree in electrical and electronics engineering from Middle East Technical University, Ankara, Turkey, in 1990, and the M.A.Sc. degree and the Ph.D. degree, both in electrical and computer engineering, from the University of Toronto, Toronto, ON, Canada, in 1992 and 1998, respectively. From 1993 to 1994, he was with the Research and Development Group, Marconi Kominikasyon A.S., Ankara. Since 1998, he has been with the Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada, where he is currently a Full Professor. From 2011 to 2012, he was with the TOBB University of Economics and Technology, Ankara, as a Visiting Professor. In recent years, his research has been funded by Huawei, Telus, Allen Vanguard, Blackberry, Samsung, Communications Research Centre of Canada, Mapsted, and DragonWave. This collaborative research resulted in over 30 patents (granted and applied). His research interests include wireless technologies with a special emphasis on wireless networks. He is a registered professional engineer in the province of Ontario, Canada. He is a Distinguished Lecturer of the IEEE Communications Society and a Distinguished Speaker of the IEEE Vehicular Technology Society. He was a recipient of the IEEE Ottawa Section Outstanding Educator Award in 2014, the Carleton University Faculty Graduate Mentoring Award in 2010, the Carleton University Graduate Students Association Excellence Award in Graduate Teaching in 2010, and the Carleton University Research Achievement Award in 2009. He has been involved in the organization of the IEEE Wireless Communications and Networking Conference (WCNC) from its inception, including serving as a Steering Committee Member and the Technical Program Chair or the Co-Chair of WCNC 2004 (Atlanta, GA, USA), WCNC 2008 (Las Vegas, NV, USA), and WCNC 2014 (Istanbul, Turkey). He was the General Co-Chair of the IEEE Vehicular Technology Conference (VTC) 2010-Fall, Ottawa. He is serving as the General Chair of the IEEE VTC 2017-Fall, which will be held in Toronto. He was the Chair of the IEEE Wireless Technical Committee. He has served on the editorial boards of the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and the IEEE COMMUNICATIONS SURVEYS AND TUTORIALS.