Queue-Aware Resource Allocation for Downlink OFDMA Cognitive Radio Networks

Patrick Mitran, *Member, IEEE*, Long Bao Le, *Member, IEEE*, Catherine Rosenberg, *Senior Member, IEEE*

***Abstract*— In this paper we consider resource allocation for an OFDMA-based cognitive radio point-to-multipoint network with fixed users. Specifically, we assume that secondary users are allowed to transmit on any subchannel provided that the interference that is created to any primary users is below a critical threshold. We focus on the downlink.**

We formulate the joint subchannel, power and rate allocation problem in the context of finite queue backlogs with a total power constraint at the base station. Thus, users with small backlogs are only allocated sufficient resources to support their backlogs while users with large backlogs share the remaining resources in a fair and efficient fashion.

Specifically, we formulate the problem as a max-min problem that is queue-aware, i.e., on a frame basis, we maximize the smallest rate of any user whose backlog cannot be fully transmitted. While the problem is a large non-linear integer program, we propose an iterative method that can solve it exactly as a sequence of linear integer programs, which provides a benchmark against which to compare fast heuristics.

We consider two classes of heuristics. The first is an adaptation of a class of multi-step heuristics that decouples the power and rate allocation problem from the subchannel allocation and is commonly found in the literature. To make this class of heuristics more efficient we propose an additional (final) step. The second is a novel approach, called selective greedy, that does not perform any decoupling. We find that while the multi-step heuristics does well in the non-cognitive setting, this is not always the case in the cognitive setting and the second heuristic shows significant improvement at reduced complexity compared to the multi-step approach.

Finally, we also study the influence of system parameters such as number of primary users and critical interference threshold on secondary network performance and provide some valuable insights on the operation of such systems.

***Index Terms*— Cognitive radio, spectrum access, resource allocation, OFDMA.**

I. Introduction

In part due to the fact that spectrum utilization in many bands is very low [1], there has recently been a large research effort in the study of secondary spectrum radio systems [3]-[7]. These systems are often called cognitive due to the sensing and advanced decision making abilities required to take advantage of licensed spectrum in a non-disruptive manner to primary users.

In this paper, we consider a downlink resource allocation (RA) problem for a point-to-multipoint cognitive wireless network. Specifically, we consider an OFDMA-based cognitive network with one base station and multiple secondary users that communicate with the base station in a single hop.

The OFDMA system consists of orthogonal subchannels, where a subchannel can be thought of as a contiguous group of subcarriers, though this is not explicitly assumed. The secondary system may transmit on any of the orthogonal subchannels provided that the interference created to a primary user, should there be one operating on a subchannel, is below a critical threshold ω chosen to guarantee that no harmful interference is created to the primary user.

We assume that perfect distributed sensing is performed by the base station and secondary users at the beginning of every frame. As a result, for each subchannel, a transmit power constraint is determined at the secondary base station that ensures that no harmful interference is created to any primary user by the secondary base station transmitting on that subchannel. These constraints are valid for the duration of the frame. We denote this collection of per subchannel transmit power constraints as vector T. In this paper we focus on studying resource allocation methods in this cognitive setting where the resources available for the secondary network evolve with time based on the activities of the primary users.

More precisely, we are interested in joint subchannel, rate and power allocation for the downlink of the secondary network. As opposed to some existing work on OFDMA resource allocation where the allocation is performed over a single time slot of a frame and then repeated for each time slot of the frame and infinite queue backlogs are usually assumed [11]-[15], we consider a general resource allocation over multiple time slots in a frame with finite queue backlog for each user to avoid over-allocation of radio resources.

Specifically, we assume that time is slotted and divided into frames of L time slots. The resource allocation (RA) problem is computed at the beginning of each frame and the corresponding resource allocation map is sent to the users so that they can tune their radio parameters to the right subchannels on a time-slot basis. The computation is done based on the channel gains measured by each user and reported to the base station, the power constraints given by vector T that depend on the activity of the primary network as well as the current queue backlogs. Hence the RA problem is clearly dynamic since from one frame to the next, new gains,

Manuscript received 16 June 2009; revised 8 Dec. 2009; accepted 15 July 2010. The associate editor coordinating the review of this paper and approving it for publication was R. Cheng. This work was presented in part at the IEEE Wireless Communications and Networking Conference (WCNC '09).

P. Mitran is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario N2L 3G1 (email: pmitran@ecemail.uwaterloo.ca).

L. B. Le is with the Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA 02139 (email: longble@mit.edu).

C. Rosenberg is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario N2L 3G1 (email: cath@ece.uwaterloo.ca).

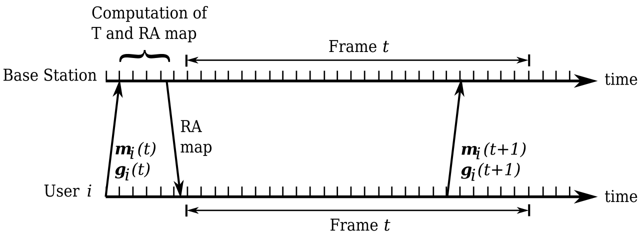

Fig. 1. Resource allocation timeline.

new power constraints and new packets arrivals will be taken into account. The evolution of user queue backlogs depends on traffic characteristics (i.e., the arrivals of packets) and on the departure of packets which depend on the available radio resources and the resource allocation strategy. We assume that the frame length L is small enough that the channel gains and the vector T remain unchanged over a frame and that the new arrivals of data packets at the base station can only be taken into account at the beginning of a frame.

Fig. 1 shows the timeline of the process including the underlying signalling protocol. Specifically prior to the beginning of frame t, each user i transmits to the base station its sensing information vector $\mathbf{m}_i(t)$ as well as its latest channel gain vector $\mathbf{g}_i(t)$ which was obtained based on pilot symbols. Based on this information and the current backlog for each user, the base station performs resource allocation for frame t. The resource allocation map is then sent to the users and is valid for the remainder of the frame. Any new packet arrivals at the base station must await the beginning of the next frame before they can be scheduled. Thus, if user i has a queue backlog $q_i(t)$ at the base station at the beginning of frame t, is allocated $x_i(t)$ packets during frame t (i.e., the base station will send $x_i(t)$ packets to i) and $a_i(t)$ new packets arrive at the base station for user i during frame t, then the new backlog at the beginning of frame $t+1$ is

$$q_i(t+1) = \max\{q_i(t) - x_i(t), 0\} + a_i(t). \tag{1}$$

The time to compute the RA solution should be significantly smaller than the duration of a frame which imposes stringent time constraints on the RA algorithm. Note that the constraint on time is critical in that the secondary network has to respond quickly to changes in primary subchannel usage to protect the primary users. This makes this problem fundamentally different from a pure OFDMA RA where insufficient responsiveness merely results in a suboptimal allocation.

To help achieve this responsiveness, instead of optimizing the resource allocation over all L time slots of the frame, an allocation over $1 \leq F \leq L$ time slots is computed and then repeated $k := L/F$ (F is assumed to divide L) times in each frame. Note that allocation over $F = 1$ time slot is less computationally heavy than over multiple time slots but the case $F > 1$ captures a practical implementation aspect since it not only improves the granularity of the resource sharing, but is necessary when the number of subchannels is smaller than the number of users or most subchannels are used by primary users with very strict power constraints on the secondary user.

The contributions of our paper are as follows:

- We formulate a resource allocation problem with finite queue backlogs over multiple time slots for the downlink of an OFDMA-based cognitive radio network. This is a non-linear problem with integer variables and thus very difficult to solve in general. We propose an iterative procedure to solve it exactly using a commercial integer program solver. Compared to much of the OFDMA RA literature where the optimal solution is rarely computed for large networks, we show that the problem can be solved exactly by a commercial solver for relatively large systems, clearly at the expense of significant computation time. This is important since it provides the optimal performance (i.e., a benchmark) against which the heuristics may be evaluated.
- On the modeling front, we have introduced the vector T that allows us to decouple the RA problem from distributed sensing and allow for interference control by the means of a critical interference threshold parameter ω.
- For online implementation, i.e., to compute the allocation in a time significantly lower than the duration of a frame, we look at two types of heuristics. The first is an adaptation of a class of decoupling heuristics common in the literature. The second is a novel multi-option greedy heuristic called selective greedy in the following. We find that the first heuristic, while it performs well in the non-cognitive setting, is clearly outperformed in the cognitive setting by the lower complexity selective greedy approach.
- On the engineering front, we find that taking queues into consideration has the potential to significantly increase the rate offered to highly backlogged users by not wasting resources on lightly-loaded users. Our study allows us to quantify this increase. We also quantify the performance improvement by performing resource allocation over multiple time slots and find it to be significant even for small values of F. Finally, we study the effect of the critical interference threshold ω to protect primary users and find that most of the gain can be achieved at surprisingly *reasonable* values.

The remainder of this paper is organized as follows. In Section II we review some related work while in Section III the system model is described and the resource allocation problem is formulated. In Section IV an exact iterative solution approach is presented and in Section V we describe the heuristics. Complexity analysis is performed in Section VI, numerical results are presented in Section VII and conclusions are stated in Section VIII.

II. Related Work

A good survey of different spectrum access models and regulatory policies can be found in [9] while [3] considers secondary spectrum access from an information theoretic point of view.

In [8] the problem of optimal channel sensing and access for opportunistic spectrum access is formulated as a partially-observable Markov decision process. In [5], the joint admission control and power allocation problem for CDMA-based spectrum sharing under the spectrum underlay paradigm is considered.

Closely related to this work is [6] where optimal power allocation for a single user under continuous rate assumption for an OFDM-based cognitive radio is handled. Our current paper considers a more general multi-user scenario with max-min rate sharing among secondary users for a downlink OFDMA-based cognitive radio network with discrete subcarrier rate assignments.

In [7], an efficient dynamic frequency hopping strategy for multi-cell IEEE 802.22 is proposed and evaluated. The proposed strategy provides a conflict-free channel allocation for 802.22-based multi-cell cognitive radio networks. In addition, it shows how out-of-band spectrum sensing can be done such that interruption of data transmission required by in-band spectrum sensing can be avoided. In [19], the limitation of the current MAC of the IEEE 802.22 standard with the hidden incumbent problem is described and solved. A distributed sensing approach is proposed in [4] and a sensing approach based on the cyclostationary properties of primary signals is presented in [18]. In [20], the performance gains due to spectrum agility, where secondary users can track available channels, are compared to the case with no agility where secondary users keep sensing and accessing a fixed channel. In [10], a physical layer implementation for an OFDMA-based cognitive radio was proposed and its performance is investigated.

Resource allocation in traditional OFDMA-based wireless networks has been an active research topic. For the downlink case, there are several important resource allocation problems. The first one is to minimize the total transmission power while providing certain required transmission rates for different users [11], [12]. The second problem optimizes a given function of the transmission rates of the different users under a total power constraint at the base station [13]-[17]. These problems are referred to respectively as margin adaptive and rate adaptive in the literature [13]. In [11], the authors propose an iterative algorithm to solve the margin adaptive problem that may not be suitable for highly dynamic wireless systems requiring fast solutions. In contrast, [12] proposes fast but suboptimal algorithms where the number of subcarriers allocated to each user is first calculated and then the subcarrier allocation for all the users is performed.

The OFDMA resource allocation problem investigated in this paper is fundamentally different from existing work in the literature in the following respects. First, except for [6] which is a single user, continuous rate allocation problem, there are extra power constraints given by the vector T as a result of distributed spectrum sensing which are not present in the existing literature. This new set of power constraints limits the transmit power on each allocated subchannel and renders many techniques unapplicable.

For example, in [12] and [21] a multi-step allocation approach is proposed that first computes the number of channels that should be allocated to each user. This computation is based on the average channel gain of each user and implicitly assumes that other than for differing subchannel gains, all subchannels are equally good, while this is clearly not the case in the cognitive setting as some subchannels may have stringent transmit power constraints while others are free of any primary user.

In Section V-A, we will adapt a common multi-step approach (see for example [14]) to the problem at hand and find by numerical computations that the adapted method can have poor performance in a cognitive setting, thus motivating the study of new methods.

In addition, we explicitly consider buffer dynamics due to finite bursty traffic patterns. While buffer dynamics have been considered before (e.g., [23]) we believe this to be the first setting in which max-min fairness and buffer dynamics are jointly considered. By considering buffer dynamics, the proposed algorithms in this paper can avoid allocating too much radio resources to lightly loaded queues as done in much existing work. In addition, we allocate resources over multiple time slots which improves the granularity of the radio resources allocation.

III. System Model and Problem Formulation

We consider an OFDMA downlink resource allocation problem with M subchannels, N secondary users and one secondary base station (referred to as the base station in the following). Any one of $\bar{z}$ transmissions modes (corresponding to a particular choice of coding and modulation schemes) can be used on any subchannel where scheme z results in rate R_z on a subchannel, i.e., R_z packets can be transmitted in one time slot over the subchannel. Without loss of generality, $0 < R_1 < R_2 < \ldots < R_{\bar{z}}$ and we denote R_1 as the lowest rate transmission mode. Finally, to employ scheme z on a subchannel requires that the Signal to Noise Ratio (SNR) on the subchannel be at least above a threshold γ_z to provide some desired block error rate.

The base station has a maximum transmit power budget of $\bar{P}_{\mathsf{max}}$ in every time slot and in the absence of primary users, can allocate any portion P_j of this power budget to subchannel j, provided $\sum_j P_j \leq \bar{P}_{\mathsf{max}}$.

Due to distributed sensing, a vector T of power constraints $\bar{P}_j$ on each subchannel is available at the base station at the beginning of a frame t (for ease of notation we omit the index t in the following), i.e., the base station must further limit $P_j \leq \bar{P}_j$ to protect primary users where $T = \{\bar{P}_j\}$ (recall that since we focus on the downlink case only the base station can transmit). $\bar{P}_j$ is a function, among other things, of the critical interference threshold ω. In the case that there is no primary user on subchannel j, then $\bar{P}_j = \infty$ as there is no primary to protect, though the sum power constraint will provide a limit to P_j.

We let g_{ij} denote the channel gain from the base station to secondary user i over subchannel j at the beginning of the frame under consideration, and $f_{ij}(z)$ be the minimum power required to transmit from the base station to user i on subchannel j using transmission mode z. $f_{ij}(z)$ is a function of the corresponding channel gain g_{ij}, the SNR threshold γ_z, the noise power at the receiver and the interference from primary users, if any, on subchannel j.

We assume that packets to be transmitted are buffered at the base station and we denote by q_i the number of packets waiting for transmission to secondary user i at the beginning of the frame. Given the backlog information, whenever possible the radio resources should be allocated to each user in such a way that the corresponding allocated aggregate (over all subchannels) rate is just sufficient to support the current backlog. A user that receives enough resources in the current frame to take care of its backlog (i.e., the base station can transmit all the queued packets of this user in the current frame) is said to be satisfied while one that is not is said to be highly backlogged.

We are interested in finding the joint subchannel, rate, and power allocation for all N secondary users which maximizes the minimum aggregated rate among highly backlogged secondary users and hence provide some form of fairness among these users. While there many different notions of fairness that can be used, here max-min resource allocation is selected because in a system with fixed users, no user should be treated differently based on its relative position to the base station. Recall that the frame length is L and for efficiency reasons the resource allocation is performed over $1 \leq F \leq L$ consecutive time slots and repeated $k = L/F$ times to fill the frame.

A resource allocation is then specified by the set of binary variables

$$\begin{aligned} S = \{s_{ijzf} \in \{0,1\} | \quad & i = 1, \ldots, N; j = 1, \ldots, M; \\ & z = 1, \ldots, \bar{z}; f = 1, \ldots, F\}, \end{aligned} \tag{2}$$

where $s_{ijzf} = 1$ iff subchannel j is allocated to user i with rate R_z in time slot f of the block. The set $\mathcal{S} \subset \{0,1\}^{N \times M \times \bar{z} \times F}$ of feasible resource allocations is given by those $S \in \mathcal{S}$ for which

$$\sum_{i=1}^{N} \sum_{z=1}^{\bar{z}} s_{ijzf} \leq 1, \quad \forall j, f \tag{3}$$

$$f_{ij}(z) s_{ijzf} \leq \bar{P}_j, \quad \forall i, j, z, f \tag{4}$$

$$\sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{z=1}^{\bar{z}} f_{ij}(z) s_{ijzf} \leq \bar{P}_{\mathsf{max}}, \quad \forall f \tag{5}$$

Eq. (3) implies that a given subchannel and time slot cannot be allocated to more than one pair (i, z). Eq. (4) ensures that the choice of coding and modulation schemes does not require a transmission power that would harm a primary user, if any. Finally, eq. (5) is the per time slot constraint on the total transmit power of the base station.

Given the set of feasible resource allocations $\mathcal{S}$, we wish to determine an allocation $S \in \mathcal{S}$ which optimizes the utility of the secondary network. In the absence of queue information, or equivalently if all users are infinitely backlogged (i.e., no users can be satisfied in a frame), the utility that we consider in this paper is max-min which provides fairness in the sense that this optimizes the smallest rate of any user. Specifically, given an allocation S, secondary user i is provided over the frame with the rate

$$x_i(S) := (L/F) \sum_{j=1}^{M} \sum_{z=1}^{\bar{z}} \sum_{f=1}^{F} R_z s_{ijzf} \tag{6}$$

over the duration of the frame and the optimal network utility under the assumption of infinite backlogs is then

$$\lambda_{\mathsf{opt}}^{\infty} = \max_{S \in \mathcal{S}} \min_{i} x_i(S). \tag{7}$$

To formulate the problem with queue backlogs (so as to avoid over-allocating resources) is somewhat less straightforward. Specifically, consider a user i who has the smallest backlog $q_i = a$ at the beginning of the frame under consideration. Then it would seem that a resource allocation that does not result in an over-allocated rate $x_i > a$ would limit the (max-min) network utility to at most $q_i = a$ which is not desirable. We would like to satisfy as many users with small backlogs as possible and make sure that those with large backlogs receive a fair share of the resources. We aim at allocating each highly backlogged user a rate which is at least as much as any satisfied users and is max-min over all unsatisfied ones. Hence, we define a max-min utility over only the unsatisfied users. We first define the set

$$\Omega(S) := \{i | x_i(S) \geq q_i\} \tag{8}$$

of users for which the allocation S satisfies their queue and $\overline{\Omega}(S)$ is the complement of $\Omega(S)$ and thus the set of users that have not had their queues satisfied. The optimal utility of the secondary network is then defined by

$$\lambda_{\mathsf{opt}} := \max_{S \in \mathcal{S}} \min_{i \in \overline{\Omega}(S)} x_i(S), \tag{9}$$

where we follow the usual mathematical convention that the $\min$ over an empty set is ∞. Thus, an optimal resource allocation will satisfy each user's queue if possible. If not, over-providing a satisfied user's queue will not provide additional utility.

The problem formulated in (9) and (3) – (5) is a *very large non-linear problem with integer variables* due to the dependency of $\overline{\Omega}$ in S. It is very general and captures several important resource allocation problems. For a traditional OFDMA resource allocation problem, constraints (4) should be removed while the case for which all users are infinitely backlogged is obtained by setting $q_i = \infty$ since then (9) degenerates to (7) as no user has its queue satisfied.

Finally, while the optimization given by (7) and (3) – (5) can be formulated as a linear integer program and thus solved by an IP solver, this is not the case for the optimization given by (9) and (3) – (5) as the set of users over which the minimization is performed depends on the choice of allocation $S \in \mathcal{S}$.

Clearly, one cannot hope to solve problem (9) and (3) – (5) exactly and fast enough (i.e., at the beginning of each frame). However, it is important to obtain exact (benchmark) results for practical scenarios (i.e., of reasonable size) so that one can i) better understand the importance of some of the parameters; ii) validate the (fast) heuristics that will be developed. Thus in Section IV, an iterative solution to numerically solve the optimization problem is presented since no commercial solver can directly solve a non-linear integer program.

Note: It may be tempting to try to solve these two problems by selecting $F = 1$ and relaxing the s_{ijzf} to real numbers in the interval $[0, 1]$. While the corresponding relaxed solution could be used to create a schedule, this schedule is not guaranteed to meet the required power constraints in each time slot, but only on average over the entire schedule. Thus, there is no assurance of protecting the primary users in each time slot. Moreover, in the case of the problem given by (9) and (3) – (5), this does not change the non-linear nature of the problem due to the utility in (9).

IV. Iterative Solution Using An Integer Program Solver

In this section, we show how the problem given by (9) and (3) – (5) can be solved exactly by solving a sequence of linear IP problems.

We start by considering a modification to the objective function in (9) as

$$\bar{\lambda}_{\mathsf{opt}} := \max_{S \in \mathcal{S}} \min_i \left[x_i(S) + \mu(x_i(S), q_i) \right], \tag{10}$$

where $\mu(x, q)$ is a function which is defined as

$$\mu(x, q) := \begin{cases} 0, & \text{if } x < q \\ \Lambda, & \text{if } x \geq q \end{cases} \tag{11}$$

where Λ is a sufficiently large number.

This transformation can be interpreted as follows. For a secondary user i such that $x_i(S) \geq q_i$, $\mu(x_i(S), q_i)$ is large enough that this secondary user will not be a bottleneck for the min operation. Therefore, the $\min$ in the objective function is only applied to secondary users with queue backlogs that are not met and the optimal resource allocation for (9) and (3) – (5) is the same as that for (10) and (3) – (5).

This transformation allows us to remove the dependency in S of the set over which the minimum is taken. However the problem is still not suitable for a linear IP solver. We now discuss how to obtain the optimal solution from an iterative procedure that invokes a linear integer program solver. This procedure works by solving a modified problem where each user is required to have either a rate λ_{new} or its queue satisfied. λ_{new} is then iteratively increased until it reaches a maximum that we denote λ^*.

```
SolveCognitiveRA

Init   : $\lambda_{\mathsf{new}} = 0$
       : $\lambda_{\mathsf{old}} = -1$
1) WHILE $\lambda_{\mathsf{new}} \neq \lambda_{\mathsf{old}}$
      a) Use a linear IP solver to find the optimal solution
         of
            $\lambda^* = \max_{S=\{s_{ijzf}\}} \min_i \{x_i(S) + \mu(\lambda_{\mathsf{new}}, q_i)\}$   (12)
            subject to
            $\sum_{i=1}^{N} \sum_{z=1}^{\bar{z}} s_{ijzf} \leq 1, \quad \forall j, f$   (13)
            $f_{ij}(z) s_{ijzf} \leq \bar{P}_j, \quad \forall i, j, z, f$   (14)
            $\sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{z=1}^{\bar{z}} f_{ij}(z) s_{ijzf} \leq \bar{P}_{\mathsf{max}}, \quad \forall f$   (15)
            $\sum_{j=1}^{M} \sum_{z=1}^{\bar{z}} \sum_{f=1}^{F} R_z s_{ijzf} \geq q_i, \quad \forall i \text{ s.t. } q_i \leq \lambda_{\mathsf{new}}$
                                                                  (16)
      b) Update: $\lambda_{\mathsf{old}} = \lambda_{\mathsf{new}}$
      c) Update: $\lambda_{\mathsf{new}} = \lambda^*$
   END WHILE
```

We now show that this iterative procedure will find the optimal solution for our resource allocation problem.

Proposition 1: The algorithm `SolveCognitiveRA` converges to an optimal solution of the resource allocation problem formulated in (10) and (3) – (5).

Proof: Since $\mu(\lambda, q)$ is non-decreasing in λ, at every iteration, λ^*, the optimal value of the objective, is non-decreasing. Since the number of subchannels and the maximum rate on each subchannel are both finite, the algorithm must converge to some value λ'.

Now, we show that the converged value λ' is an optimal solution for the problem formulated in (9) and (3) – (5).

Let λ_{opt} be the optimal objective value of (9) and (3) – (5). Also, suppose that the algorithm converges to $\lambda' < \lambda_{\mathsf{opt}}$. This means that substituting either $\lambda_{\mathsf{new}} = \lambda^{'}$ or $\lambda_{\mathsf{new}} = \lambda_{\mathsf{opt}}$ into (12)–(16) yields a feasible solution. However, since $\lambda_{\mathsf{opt}} > \lambda'$ is the optimal solution, there is an allocation such that any secondary user i with $q_i \leq \lambda'$ will receive a rate at least equal to its queue backlog while other secondary users (i.e., those with $q_i > \lambda'$) can be supported at rates strictly larger than λ'. This is a contradiction because the solution in the last iteration of the algorithm provides rates of at most λ' for a secondary user i with $q_i > \lambda'$. Hence, the iterative algorithm must converge to the optimal solution. ■

V. Heuristics

In this section, we consider two heuristics for the resource allocation problem. The first is an adaptation of the multi-step decoupling heuristic of [14]. While the approach works well in the absence of primary users, the adaptation to the cognitive case does not work well in the presence of a large number of primary users in spite of adding an additional step. Thus, we also consider a second heuristic which does not decouple the allocation in sub-problems. This heuristic is called selective greedy.

A. Multi-Step Approach

The multi-step heuristic has four steps. The first three are adapted from the three proposed in [14] and the last one is novel and is called the perturbation step in the following. The details can be found in [22] and we restrict the description in this paper to the broad concepts.

Specifically, in Step 1, we perform power allocation over the subchannels by sharing $\bar{P}_{\mathsf{max}}$ as uniformly as possible considering the power constraints in vector T (i.e., there is no point allocating more than $\bar{P}_j$ to subchannel j). This results in a subchannel being allocated either a power $P_j = \bar{P}_j$ or the same power as any subchannel which is not at its limit $\bar{P}_j$. This power allocation is used in each of the F time slots.

With the power allocation of Step 1 now fixed, we perform subchannel-time slot pair allocation in Step 2. Specifically, subchannel-time slot pairs are allocated to the secondary users sequentially where in each allocation iteration, a secondary user with the smallest rate is allocated one available pair achieving the highest rate subject to the power allocation in Step 1. Ties in the subchannel-time slot pairs are broken in favor of a pair with the highest channel gain.

When a secondary user i has received enough resources to satisfy its queue, i.e., it has received an allocation of at least q_i packets, it is removed from the list of non-satisfied secondary users, and thus not allocated any more resources. If the "best" subchannel-time slot pair in any allocation iteration does not improve the rate of the secondary user under consideration, then we cannot improve the utility function and we allocate all remaining available pairs to secondary users whose queues are not yet fully satisfied in a round robin fashion in preparation for the next step. Finally, the power allocated to a subchannel in Step 1 is usually larger than the power required to deliver the assigned rate once it has been allocated to a secondary user. Therefore, after each subchannel-time slot allocation iteration, the residual power on the selected pair is calculated and allocated to the set of remaining subchannels in the time slot as evenly as possible considering the power limits due to T.

Given the subchannel allocation solution from Step 2, there is a potential max-min rate improvement by redoing rate and power allocation. This is done in Step 3. Specifically, we sequentially increment the transmission mode of the most power-efficient subchannel-time slot pair that would not violate the primary protection constraint $\bar{P}_j$ for the current minimum rate secondary user in each rate update operation. This is an adapted version of the multi-user bit-loading algorithm. In addition, as soon as the rate of a secondary user becomes greater than or equal to its queue backlog, the secondary user is removed from the list of active secondaries for all subsequent rate updates.

At this point, the allocation is feasible and if we stop the heuristic here, we denote it by Step 3.

When Step 3 terminates, the total consumed power in any of the F time slots may be still well below the maximum power budget of the base station (i.e., $P < \bar{P}_{\mathsf{max}}$). To exploit the remaining base station power, in Step 4 we perform limited perturbation on the subchannel allocation to improve the minimum rate among all secondaries whose queue is not satisfied. Specifically, for each bottleneck secondary user (i.e., a secondary user which among all those whose queue is not satisfied, has a minimum rate), we attempt to take one subchannel-time slot pair from a non-bottleneck secondary user, allocate it to the bottleneck secondary user and use single-user bit-loading to calculate the rates for both bottleneck and granting secondaries. The subchannel-time slot reassignment is only performed if the rate of the bottleneck secondary user is improved while not reducing the rate of the granting secondary user below or at the former rate of the bottleneck secondary.

Finally, if the perturbation at the end of Step 4 is successful in increasing the rate of *all* the bottleneck secondaries from Step 3, we can proceed to a new round of rate and power allocation (i.e., a new Step 3) followed by perturbation (i.e., a new Step 4). This may be repeated a fixed number of times, or until perturbation fails to increase the rate of all the bottleneck secondaries, though in practice, beyond two iterations the perturbation step was never observed to produce additional gain.

B. Selective Greedy Approach

A major drawback of the multi-step approach is that due to the decoupling of the subchannel-time slot allocation from rate and power allocation, the overall allocation may be far from optimal while the final perturbation step has limited ability to rectify the allocation.

The broad structure of the selective greedy approach is that at every iteration we attempt to increase a lowest rate secondary user whose queue is not satisfied and do so in the most efficient manner where efficiency is based on the ratio of power increase to rate increase. Three potential methods of increasing a user's rate are computed and the most efficient as measured by this ratio is selected. In each of the three methods (the methods are described below), if it is not feasible to increase the user's rate, then the power cost is said to be infinite which effectively eliminates the option.

The first method, called `NewChannel`, is to simply assign to the user i under consideration a subchannel-time slot pair among all the free subchannel-time slot pairs with the lowest rate modulation scheme R_1 (provided this does not violate the primary protection constraint $\bar{P}_j$) at the lowest power cost. Hence the input to this module is the user index i and the outputs are the power increase $\Delta\mathbf{P}_1$, the rate increase ΔR_1 and a list RA_1 comprising the (single) assigned subchannel-time slot pair (f^*, j^*).

Note that $\Delta\mathbf{P}_1$ is a vector of length F and $(\Delta\mathbf{P}_1)_\ell$ denotes its ℓth component.

```
NewChannel(user $i$)
  Init:  Let $\Delta\mathbf{P}_1 = 0 \in \mathbb{R}^F$.
         Let $\delta P_{(f,j)} = \infty \quad \forall f, j$
  1) FOR each yet unallocated subchannel-time slot pair $(f, j)$
       a) IF lowest rate on $(f, j)$ for user $i$ does not violate $\bar{P}_j$
          THEN $\delta P_{(f,j)}$ = power for lowest rate on subchannel $j$ for user $i$.
  2) END FOR
  3) Let $(f^*, j^*) = \arg\min \delta_{(f,j)}$.
  4) Let $(\Delta\mathbf{P}_1)_{f^*} = \delta P_{(f^*,j^*)}$, $\Delta R_1 = R_1$.
```

```
5) Let $RA_1 = \{(f^*, j^*)\}$.
```

The second method, called `IncrementChannel`, is to increment the rate of an already allocated subchannel-time slot pair to the secondary user i under consideration, again provided this does not violate the primary protection constraint $\bar{P}_j$. For each subchannel-time slot pair already allocated to i, we compute the lowest power necessary to increase the rate to the next lowest value, and among all these potential solutions we select the one that has the best efficiency (as defined above). Hence the input to this module is the user index i and the outputs are the power increase $\Delta\mathbf{P}_2$, the rate increase ΔR_2, and a list RA_2 comprising the (single) selected subchannel-time slot pair (f^*, j^*) whose rate is to be incremented. Note that a division by 0 in line 3 is treated as ∞.

```
IncrementChannel(user $i$)
  Init: Let $\Delta\mathbf{P}_2 = 0 \in \mathbb{R}^F$.
        Let $\delta P_{(f,j)} = \infty \quad \forall f, j$
        Let $\delta R_{(f,j)} = 0 \quad \forall f, j$
  1) FOR each allocated subchannel-time slot pair $(f, j)$ to user $i$
       a) IF rate on $(f, j)$ is less than $R_{\bar{z}}$, consider the lowest
          non-zero rate increment $\delta R$. It requires a power
          increase of $\delta P$. If this power increase does not
          violate $\bar{P}_j$
          THEN $(\delta P_{(f,j)}, \delta R_{(f,j)}) = (\delta P, \delta R)$
  2) END FOR
  3) Let $(f^*, j^*) = \arg\min \delta P_{(f,j)} / \delta R_{(f,j)}$.
  4) Let $(\Delta\mathbf{P}_2)_{f^*} = \delta P_{(f^*,j^*)}$, $\Delta R_2 = \delta R_{(f^*,j^*)}$.
  5) Let $RA_2 = \{(f^*, j^*)\}$.
```

The third method, called `SwapChannel`, is for the secondary user i under consideration to take a subchannel already allocated to another user, say i', and to use the lowest modulation scheme on this subchannel. Since the secondary user i' from which the subchannel was taken has now had its rate decreased, we must compensate by increasing the rate on its remaining assigned subchannels by assigning more power. Hence the input to this module is the user i and the outputs are the power increase $\Delta\mathbf{P}_3$, the rate increase ΔR_3 and a list RA_3 comprising multiple subchannel-time slot pairs, where the first pair is the one donated to user i (its rate being R_1), and the other pairs are those whose rate must be incremented by the smallest amount compared to their current setting to compensate the donating user i'. A subchannel-time slot pair that appears twice in the increment list sees its rate incremented twice.

Here $\Delta\mathbf{P}_{(f,j)}$ is a vector of length F that tracks the power required by the method in each of the F time slots if subchannel-time slot pair (f, j) is given to user i. The notation $(\Delta\mathbf{P}_{(f,j)})_\ell$ denotes the ℓth component of the vector of length F.

```
SwapChannel(user $i$)
  Init: Let $\Delta\mathbf{P}_{(f,j)} = \infty \in \mathbb{R}^F \quad \forall f, j$
        Let $RA(f, j)$ be empty lists of allocations for all $f, j$.
  1) Save current resource allocation in variable $CRA$.
  2) FOR each allocated subchannel-time slot pair $(f, j)$ to a user $i' \neq i$
       a) Restore resource allocation $CRA$.
       b) Let $r$ be the allocated rate of user $i'$
       c) IF lowest rate on $(f, j)$ for user $i$ violates $\bar{P}_j$ then
          set $\Delta\mathbf{P}_{(f,j)} = \infty \in \mathbb{R}^F$.
       d) ELSE $\Delta\mathbf{P}_{(f,j)}$ = change in power (in each time
          slot) due to reassignment of $(f, j)$ to user $i$.
       e) WHILE rate of user $i'$ is less than $r$ and
          $\Sigma_\ell (\Delta\mathbf{P}_{(f,j)})_\ell < \infty$
            i) $(\Delta\mathbf{P}, \Delta R_{\mathsf{inc}}, RA_{\mathsf{inc}}) =$
               IncrementChannel($i'$)
           ii) Update rate of user $i'$ and resource allocation
               based on $RA_{\mathsf{inc}}$.
          iii) $\Delta\mathbf{P}_{(f,j)} = \Delta\mathbf{P}_{(f,j)} + \Delta\mathbf{P}$
           iv) $RA(f, j) = RA(f, j) + RA_{\mathsf{inc}}$.
       f) END WHILE
  3) END FOR
  4) Let $(f^*, j^*) = \arg\min \Sigma_\ell (\Delta\mathbf{P}_{(f,j)})_\ell$.
  5) Let $\Delta\mathbf{P}_3 = \Delta\mathbf{P}_{(f^*,j^*)}$, $\Delta R_3 = R_1$.
  6) Let $RA_3 = \{(f^*, j^*)\} + RA(f^*, j^*)$.
```

The iterative selective greedy algorithm which sequentially increments the user's rates is as follows. In a given iteration, the method with the highest efficiency that does not violate the per time slot **sum** power constraint is selected. It stops when no finite ratios are found.

```
SelectiveGreedyRA($F$)
  Init: Let $\mathbf{P} = 0 \in \mathbb{R}^F$.
        Let $\Delta\mathbf{P}_k \in \mathbb{R}^F$, $k = 1, 2, 3$.
        Let $\Delta R_k \in \mathbb{R}$, $k = 1, 2, 3$.
        Let $RA_1$, $RA_2$ and $RA_3$ be the output lists of the
        three methods.
        Let $bContinue = true$.
  1) REPEAT
       a) Find user $i$ with unsatisfied queue and lowest rate.
       b) Let $(\Delta\mathbf{P}_1, \Delta R_1, RA_1) =$ NewChannel($i$).
       c) Let $(\Delta\mathbf{P}_2, \Delta R_2, RA_2) =$
          IncrementChannel($i$).
       d) Let $(\Delta\mathbf{P}_3, \Delta R_3, RA_3) =$ SwapChannel($i$).
       e) Let $\xi_k = \sum_\ell (\Delta\mathbf{P}_k)_\ell / \Delta R_k$, for $k = 1, 2, 3$.
       f) Let $T$ be the set of $k$ such that a change in
          power $\Delta\mathbf{P}_k$ would not violate the per time-slot
          sum power constraints.
       g) Let $k^* = \arg\min_{k \in T} \xi_k$.
       h) IF $\xi_{k^*} < \infty$, then apply allocations in list $RA_{k^*}$.
       i) ELSE $bContinue = false$.
  2) WHILE $bContinue = true$.
```

VI. Complexity

We now analyze the complexity of the heuristics in terms of the number of operations over F time slots. The analysis does not depend on the particular choice of queue lengths other than the assumption that there is always at least one user whose queue cannot be satisfied.

A. Multi-step Approach

In Step 1, we have to repeatedly search for the smallest $\bar{P}_j$ among all subchannels that are not yet allocated power at $\bar{P}_j$. Since there are M subchannels, this requires at most $O(M^2)$ operations

In Step 2, before each subchannel allocation, the minimum-rate secondary user to which the allocation is performed must be found. This requires $O(N)$ operations.

Given the selected secondary user, the subchannel achieving maximum rate with least power will be chosen for allocation. In fact, in the k-th allocation iteration, there remains $F \times M - k + 1$ subchannels-time slot pairs. For each of these remaining pairs, at most $\bar{z}$ comparisons are needed to find the maximum achievable rate. The total number of comparisons needed for this operation is

$$\sum_{k=1}^{FM} (FM - k + 1)\bar{z} = O(\bar{z}F^2M^2). \tag{17}$$

In the worst case where $FM - k + 1$ available subchannels achieve the same rate, $FM - k$ subchannel gain comparisons are needed to find the one with least power. Therefore, the total number of operations in the worst case is

$$\sum_{k=1}^{FM} (FM - k + 1)(FM - k) = O(F^3M^3) \tag{18}$$

Now, we analyze the complexity due to the redistribution of residual power in each subchannel allocation iteration. For the k-th allocation from a time slot, there are at most $M - k$ subchannels to receive residual power. In the worst case, the number of operations to perform is $(M - k)(M - k - 1)/2$. Therefore, in the worst case, the number of operations needed for the residual power allocation step is

$$F \sum_{k=1}^{M-1} \frac{(M - k)(M - k - 1)}{2} \leq O(FM^3). \tag{19}$$

Since in practice $\bar{z} << M$, the complexity for this step is $O(F^3M^3)$.

Let C_i be the set of channel-time slot pairs allocated to user i after Step 2. In Step 3, for secondary user i there are at most $\bar{z}\,|C_i|$ rate updates for all the allocated subchannels. In each rate update for secondary user i, there are at most $|C_i|$ feasible subchannels to choose. In the worst case, the number of comparisons needed to find the most power-efficient subchannel is $|C_i| - 1$. Therefore, the maximum number of comparisons needed in the worst case for secondary user i is $\bar{z}\,|C_i|\,(|C_i| - 1)$. Considering all the secondary users, the worst-case complexity occurs when only one secondary user is allocated all the subchannels. In this case, the total number of comparisons required in Step 3 is

$$\sum_{i=1}^{N} \bar{z}\,|C_i|\,(|C_i| - 1) = O(\bar{z}F^2M^2). \tag{20}$$

We now analyze the complexity of one subchannel perturbation per bottleneck secondary user in Step 4. Consider one subchannel reassignment from secondary i to secondary $\bar{i}$. There are $|C_i|$ possible ways to choose one subchannel from secondary i. For each possible subchannel reassignment, we have to redo rate and power allocation for secondary $\bar{i}$ which requires up to $\bar{z}\,|C_i|\,(|C_i| - 1)$ operations and $\bar{z}\,|C_{\bar{i}}|\,(|C_{\bar{i}}| - 1)$ for secondary $\bar{i}$. The worst case complexity in the number of operations is

$$\bar{z}\,|C_i|\,\{|C_i|\,(|C_i| - 1) + |C_{\bar{i}}|\,(|C_{\bar{i}}| - 1)\}\,. \tag{21}$$

Let $\mathcal{B}$ be the set of bottleneck secondary users and $\bar{\mathcal{B}}$ be the complement of $\mathcal{B}$. In the worst case we have to try all possible secondary users in set $\bar{\mathcal{B}}$ to find one granting secondary for each bottleneck secondary $\bar{i} \in \mathcal{B}$. Hence, the worst case complexity order for one iteration of perturbations in Step 4 in terms of operations is

$$\sum_{j \in \mathcal{B}} \sum_{\bar{i} \in \bar{\mathcal{B}}} \bar{z}\,|C_i|\,\{|C_i|\,(|C_i| - 1) + |C_{\bar{i}}|\,(|C_{\bar{i}}| - 1)\}$$
$$\leq 2\bar{z}F^3M^3N = O(\bar{z}F^3M^3N). \tag{22}$$

We therefore find that the complexity of Steps 1-3 is $O(F^3M^3)$ while that of Steps 1-4 is dominated by the last step with a per iteration complexity of $O(\bar{z}F^3M^3N)$. It should be noted that in practice the perturbation step was never applied more than two times before it was unable to provide rate improvement.

B. Selective Greedy Approach

We now analyze the complexity of the selective greedy scheme in terms of the number of operations required.

First, we observe that in every loop that `CognitiveIncrementalRA` executes, one user has its rate incremented. Thus, `CognitiveIncrementalRA` may loop no more than $\bar{z}FM$ times.

Second, in each loop, a lowest rate user must be found and 3 possible incremental allocations methods are evaluated. A lowest rate user can be found in $O(N)$ operations.

The first method attempts to assign an unallocated subchannel-time slot to the user. Since there are at most $F \times M$ such subchannels, this method has a worst case complexity $O(FM)$.

The second method attempts to increment the rate of an already assigned subchannel to the user. Again, since there are at most $F \times M$ such subchannels, this method has a worst case complexity $O(FM)$.

The third method attempts to re-assign a subchannel from another user to the user under consideration and then increment the rates on the donating user's allocated subchannel-time slot pairs by running the second method as needed. There are at most $F \times M$ time slot-channel pairs that can be donated and the number of times that the second method is run is at most $R_{\bar{z}}/R_1$. If the rates are taken to be sequential integers as in the next section, then $R_{\bar{z}}/R_1 = \bar{z}$ and the complexity of the third method is at most $O(\bar{z}F^2M^2)$.

Combining these, and assuming that $\bar{z}F^2M^2 >> N$ we find that the complexity of the entire scheme is at most $O(\bar{z}^2F^3M^3)$.

This compares favorably with the multi-step scheme which has a complexity of $O(\bar{z}F^3M^3N)$ since in practice $\bar{z} << N$.

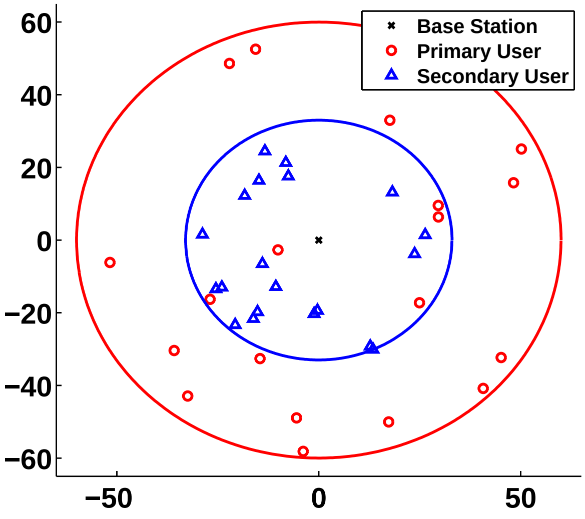

Fig. 2. Sample placement of 20 primary and 20 secondary users.

VII. Numerical Results

A. Input Generation

To evaluate the performance of the proposed heuristics, we now describe how we generate realistic values for the vector T and the $f_{ij}(z)$. The channel gain between the base station and a receiver i (primary or secondary) at distance d_i from the base station on subchannel j is modeled as a combination of path loss and fading. In particular, we model the received power, P_R, by $P_R = g_{ij} P_T$ where P_T is the transmit power and

$$g_{ij} = |h_{ij}|^2 \left(d_0/d_i\right)^\eta , \tag{23}$$

where h_{ij} is an independent Ricean fading gain characterized by its K-factor, η is the path loss exponent and d_0 the far-field crossover distance.

We generate randomly and uniformly the positions of N secondary users in a disk of radius r_2 centered on the base station while N_p primary receivers are placed uniformly and randomly in a disk of radius $r_1 > r_2$ centered on the base station. We model the primary channel occupancy by randomly and uniformly assigning one subchannel to each primary receiver such that no two primary users occupy the same subchannel. Denote the subchannel for primary receiver n by j_n. Then, the $\bar{P}_j$ are taken to be the largest feasible value such that the received power from the base station to the primary user n on channel j_n is at most ωN_0 where N_0 is the Gaussian noise power and ω is the critical interference threshold which allows one to adjust the maximum amount of interference that can be caused to a primary user. For simplicity we take $f_{ij}(z) = \gamma_z N_0 / g_{ij}$ though we could incorporate primary interference to secondary users in a more complex model.

Using the method described above, all numerical results that were generated share the following parameters. The K-factor is -10 dB which reflects scenarios with little to no line of sight, $\eta = 3$, $d_0 = 50$ m, $r_1 = 33$ km, $r_2 = 60$ km, and $N_0 = -100$ dB. Thus a device at a distance of 33 km from the base station will see an average (neglecting fading) SNR of 15.4 dB if the base station uses a transmit power of 1 Watt. We use in all cases five transmission modes of rates $R_1 = 1$, $R_2 = 2$, $R_3 = 3$, $R_4 = 4$ and $R_5 = 5$ with SINR thresholds of $\gamma_1 = 10$ dB, $\gamma_2 = 14.77$ dB, $\gamma_3 = 18.45$ dB, $\gamma_4 = 21.76$ dB and $\gamma_5 = 24.91$ dB. Unless otherwise stated, the critical interference threshold $\omega = 0$ dB which corresponds to allowing the secondary network to create an interference to the primary users of at most the same level as the noise power.

To obtain the average max-min rate for a given scenario characterized by (M, N, N_p), we average the corresponding results over 20 independent generations of node positions and fading coefficients. In the test cases below, the number of secondary users is chosen to be 40. Finite queue backlogs at the beginning of a frame, if applicable, are selected such that 5 users each have queue backlogs of 90, 180, 270 and 360 while the remaining 20 users have backlogs of 900. We choose $L = 30$ and consider $F = 1$ and $F = 3$. In the remainder we will give results in terms of per frame performance.

B. Discussion

We start by considering the optimal theoretical performance of the system given by (9) (i.e., λ_{opt}) obtained by using the iterative solution described in Section IV where we have used CPLEX, a commercial integer program solver, in each step to solve the linear IP. Note that in the figures, the label “No Queue” indicates that the algorithm assumes that the users are greedy while the label “Queues” indicates that the algorithms take the value of the queue backlogs into account.

Fig. 3 shows λ_{opt} averaged over the 20 realizations as a function of P_{max} for $F = 1$ and $F = 3$, with and without queues and without any primary users for $M = 120$ subchannels and $N = 40$ while Fig. 4 considers the case of $N_p = 60$ primary users, i.e., half of the 120 subchannels are occupied by primaries.

Both figures show that increasing the power budget increases the max-min rate but with diminishing returns and quantify how much can be gained by taking queues into account. Clearly taking queue backlogs into consideration can greatly improve performance.

¿From Fig. 3 we see that for a traditional allocation problem (i.e., without primary users) and in the absence of queues, there is little gain in principle in performing max-min rate allocation over multiple time slots (in this case $F = 3$) as opposed to a single time slot ($F = 1$). When queues are taken into account, this is no longer the case. Specifically, we see that for the parameters chosen, an average gap of 8% is noticed at the maximum power setting, though in a more power limited regime the gap is less.

¿From Fig. 4 we see that in the presence of primary users, this is no longer the case. Specifically in the case that there are no queues and $F = 1$, we see that the max-min rate of the users has saturated while this is not the case when $F = 3$. We attribute this to the fact that resource allocation over multiple time slots provides better granularity and thus is better able to exploit subchannels occupied by primary users since these subchannels can only be effectively used by users who are near the base station due to the constraints $\bar{P}_j$.

In Fig. 5, the average max-min rate is shown versus the power constraint P_{max} for the three choices of the interference power control parameter ω = -20dB, 0dB and 10dB. The selection of $\omega = -20$dB results in interference to primary users that is at most $N_0/100$ and thus, has negligible impact on the primary SINR, i.e., secondary users essentially avoid creating interference to primary users. The selection $\omega = 0$dB result in interference that is at most N_0, and thus, the SINR of primary users decreases by at most 3dB, while the selection $\omega = 10$dB decreases the SINR by approximately 10.4dB. Interestingly, we find that the selection $\omega = 0$dB results in significant gain compared to $\omega = -20$ and in fact increasing to $\omega = 10$dB improves the average max-min rate by only 15% in this case.

In Fig. 6, the average max-min rates of the users is shown in the absence of primary users for optimal allocation as well as the multi-step heuristic after Step 3 and Step 4 and the selective greedy heuristic. We find that Step 3 of the multi-step heuristic significantly under performs compared to the optimal performance. By comparison, both the results of Step 4 as well as the selective greedy heuristic are nearly optimal, both with and without queues, though the selective greedy heuristic has lower complexity.

In Fig. 7, the average max-min rates of the users is shown in the presence of 60 primary users for optimal allocation as well as the multi-step heuristic after Step 3 and Step 4 and the selective greedy heuristic. Here, we see that without taking queues into account, the selective greedy heuristic as well as Step 4 perform well while Step 3 shows a significant gap. Interestingly, when queues are taken into account, none of the heuristics performs close to optimal, though the selective greedy approach shows the best performance.

Finally, in Fig. 8, we consider the case of Fig. 7, though with $F = 3$. Here, we find that even though the optimal performance has significantly increased, the selective greedy heuristic has significantly narrowed the gap to the optimal performance. We attribute this to the better granularity afforded by the larger F. By comparison, the outcome of Step 4 shows a significant gap.

VIII. Conclusions

We have considered OFDMA cognitive resource allocation for downlink cognitive radio networks. We have proposed an iterative procedure involving a sequence of integer linear programs to solve this non-linear integer problem and found that the optimal performance could be computed for relatively large problems.

We have proposed two heuristics. The first is an adaptation of a class of decoupling algorithms common in the literature. The second does not decouple the problem and we refer to it as selective greedy. The second heuristic clearly outperforms the first in the cognitive setting.

This paper has numerically quantified the performance gain obtained by taking queues into account and the improvement obtained by allocating resources over multiple time slots. Finally, we have found that surprisingly good performance can be obtained at reasonable values of the interference control parameter ω. Specifically, $\omega = 0$ dB yields most of the gain that is to be had.

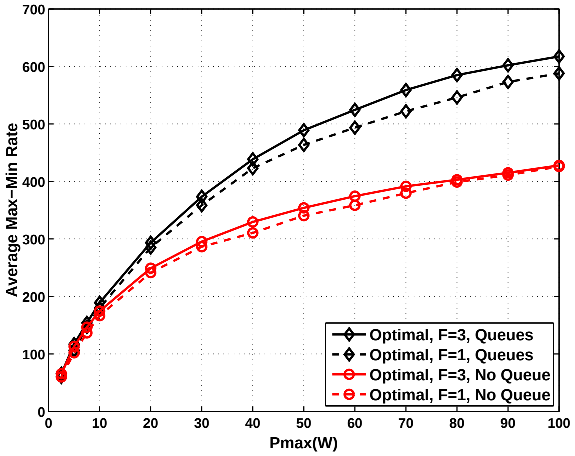

Fig. 3. Average per frame max-min rate of the optimal solution versus $\bar{P}_{\mathsf{max}}$ for $F = 1, 3$, $(M, N, N_p) = (120, 40, 0)$

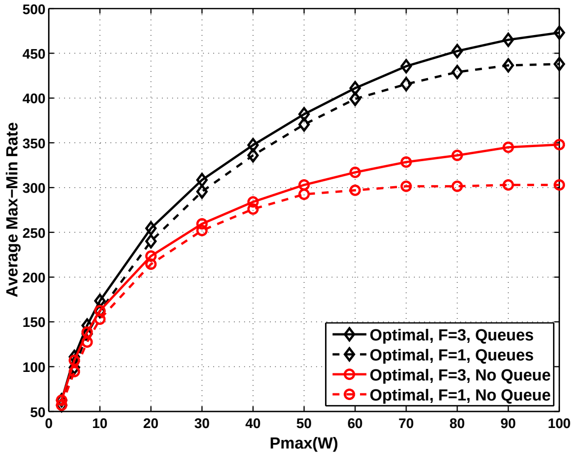

Fig. 4. Average per frame max-min rate of the optimal solution versus $\bar{P}_{\mathsf{max}}$

References

[1] FCC Spectrum Policy Task Force, “Report of the spectrum efficiency working group,” FCC, Tech. Rep., Nov. 2002.

[2] C. Cordeiro, K. Challapali, D. Birru, and N. S. Shankar, “IEEE 802.22: The first worldwide wireless standard based on cognitive radios,” in Proc. *DySpan 2005*, Nov. 2005.

[3] N. Devroye, P. Mitran and V. Tarokh, “Achievable rates in cognitive radio channels,” *IEEE Trans. Inf. Theory*, vol. 52, no. 5, pp. 1813-1827, 2006.

[4] M. Gandetto and C. Regazzoni, “Spectrum sensing: A distributed approach for cognitive terminals,” *IEEE J. Sel. Areas Commun.*, vol. 25, no. 3, pp. 546-557, April 2007.

[5] L. Le and E. Hossain, “QoS-aware spectrum sharing in cognitive wireless networks,” in Proc. *IEEE GLOBECOM’2007*, Nov. 2007.

[6] P. Wang, M. Zhao, L. Xiao, S. Zhou, and J. Wang, “Power allocation in OFDM-based cognitive radio systems,” in *IEEE GLOBECOM*, 2007.

[7] H. Wendong, D. Willkomm, M. Abusubaih, J. Gross, G. Vlantis, M. Gerla, and A. Wolisz, “Dynamic frequency hopping communities for efficient IEEE 802.22 operation,” *IEEE Commun. Mag.*, pp. 80-87, vol. 45, no. 5, May 2007.

[8] Q. Zhao, L. Tong, A. Swami, and Y. Chen “Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: A POMDP framework,” *IEEE J. Sel. Areas Commun.*, vol. 25, no. 3, pp. 589-600, April 2007.

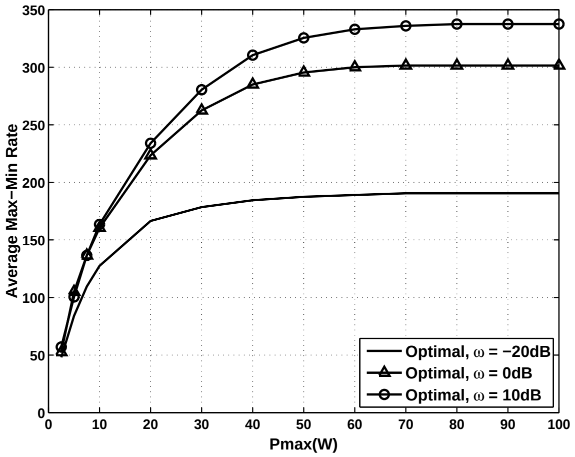

Fig. 5. Average max-min per frame rate of the optimal solution versus $\bar{P}_{\text{max}}$ for $F = 1$, $(M, N, N_p) = (120, 40, 50)$ for several critical interference threshold ω choices.

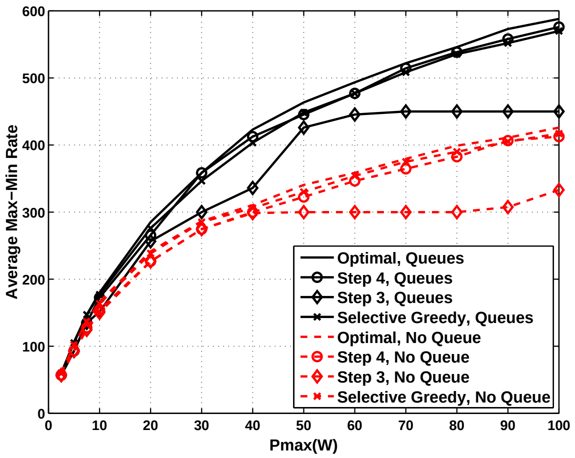

Fig. 6. Average max-min per frame rate of the optimal solution and heuristics versus $\bar{P}_{\text{max}}$ for $F = 1$, $(M, N, N_p) = (120, 40, 0)$.

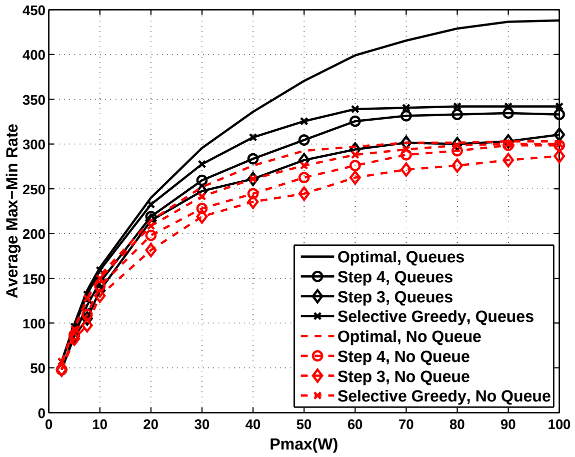

Fig. 7. Average max-min per frame rate of the optimal solution and heuristics versus $\bar{P}_{\text{max}}$ for $F = 1$, $(M, N, N_p) = (120, 40, 60)$.

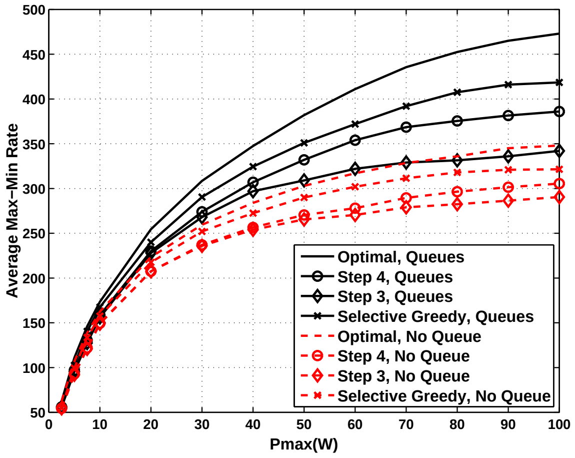

Fig. 8. Average max-min per frame rate of the optimal solution and heuristics versus $\bar{P}_{\text{max}}$ for $F = 3$, $(M, N, N_p) = (120, 40, 60)$.

[9] Q. Zhao and B. M. Sadler, “A survey of dynamic spectrum access: Signal processing, networking, and regulatory policy,” *IEEE Signal Processing Mag.*, pp. 79-89, May 2007.

[10] R. Rajbanshi, A. M. Wyglinski, and G. J Minden, “An efficient implementation of NC-OFDM transceivers for cognitive radios,” in *Proc. CROWNCOM 2006*, June 2006.

[11] C. Y. Wong, R. S. Cheng, and K. B. Letaief, and R. D. Murch, “Multiuser OFDM with adaptive subcarrier, bit, and power allocation,” *IEEE J. Sel. Areas Commun.*, vol. 17, no. 10, pp. 1743-1758, Oct. 1999.

[12] D. Kivanc, G. Li, and H. Liu, “Computationally efficient bandwidth allocation and power control for OFDMA,” *IEEE Trans. Wireless Commun.*, vol. 2, no. 6, pp. 1150-1158, Nov. 2003.

[13] I. Kim, I. S. Park, and Y. H. Lee, “Use of linear programming for dynamic subcarrier and bit allocation in multiuser OFDM,” *IEEE Trans. Veh. Technol.*, vol. 55, no. 4, pp. 1195-1207, July 2006.

[14] W. Rhee and J. M. Cioffi, “Increase in capacity of multiuser OFDM system using dynamic subchannel allocation,” in *Proc. IEEE VTC’2000.*

[15] C. Zeng, L. M. C. Hoo, and J. M. Cioffi, “Efficient water-filling algorithms for a Gaussian multiaccess channel with ISI,” in Proc. *IEEE VTC’2000*, Sept. 2000.

[16] J. Jang and K. B. Lee, “Transmit Power Adaptation for Multiuser OFDM Systems” *IEEE J. Sel. Areas Commun.*, vol. 21, no. 2, pp. 171-178, Feb. 2003.

[17] G. Song and Y. Li, “Cross-layer optimization for OFDM wireless networks, Part II: Algorithm development,” *IEEE Trans. Wireless Commun.*, vol. 4, no. 2, pp. 625-634, Mar. 2005.

[18] H. -S. Chen, W. Gao, and D. G. Daut, “Spectrum sensing using cyclostationary properties and application to IEEE 802.22 WRAN,” in *IEEE GLOBECOM 2007*, Nov. 2007.

[19] S. Sengupta, S. Brahma, M. Chatterjee, and N. S. Shankar, “Enhancements to cognitive radio based IEEE 802.22 air-interface,” in *Proc. IEEE ICC 2007*, May 2007.

[20] C. T. Chou, S. Shankar. N, H. Kim, and K. G. Shin, “What and how much to gain by spectrum agile?” *IEEE J. Sel. Areas Commun.*, vol. 25, no. 3, pp. 576-588, April 2007.

[21] H. Yin and H. Liu, “An Efficient Multiuser Loading Algorithm for OFDM-based Broadband Wireless Systems,” in *Proc. IEEE Globecom 2000*, Nov. 2000.

[22] L. Le, P. Mitran and C. Rosenberg, “Queue-Aware Subchannel and Power Allocation for Downlink OFDM-Based Cognitive Radio Networks,” in *Proc. IEEE WCNC*, Apr. 2009.

[23] S. Kittipiyakul and T. Javidi, “Delay-Optimal Server Allocation in Multiqueue Multiserver Systems With Time-Varying Connectivities,” *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2319 – 2333, May 2009.

Patrick Mitran (S'01, M'07) received the Bachelor's and Master's degrees in electrical engineering, in 2001 and 2002, respectively, from McGill University, Montreal, PQ, Canada, and the Ph.D. degree from the Division of Engineering and Applied Sciences, Harvard University, Cambridge, MA in 2006. In 2005, he interned as a research scientist for Intel Corporation in the Radio Communications Lab. In 2006-07 he was an applied mathematics lecturer in the School of Engineering and Applied Sciences, Harvard University. Since fall 2007, he is with the Department of Electrical and Computer Engineering at the University of Waterloo at the rank of Assistant Professor.

P. Mitran has interests in wireless networks, communication theory, signal processing, coding theory and information theory. Some of his most notable contributions are in cognitive radio, cooperative communications, distributed beamforming and distributed data compression.

Long Bao Le (S'04-M'07) received the B.Eng. degree from Ho Chi Minh City University of Technology, Vietnam, in 1999, the M.Eng. degree from Asian Institute of Technology (AIT), Thailand, in 2002 and the Ph.D. degree from University of Manitoba, Canada, in 2007. He is currently a Postdoctoral Research Associate at Massachusetts Institute of Technology, USA. His current research interests include cognitive radio, link and transport layer protocol issues, cooperative diversity and relay networks, stochastic control and cross-layer design for communication networks.

Catherine Rosenberg is a Professor in Electrical and Computer Engineering at the University of Waterloo. Since June 2010, she holds the Canada Research Chair in the Future Internet. She started her career in ALCATEL, France and then at AT&T Bell Labs., USA. From 1988-1996, she was a faculty member at the Department of Electrical and Computer Engineering, Ecole Polytechnique, Montral, Canada. In 1996, she joined Nortel Networks in the UK where she created and headed the R&D Department in Broadband Satellite Networking. In August 1999, Dr. Rosenberg became a Professor in the School of Electrical and Computer Engineering at Purdue University where she co-founded in May 2002 the Center for Wireless Systems and Applications (CWSA). She joined University of Waterloo on Sept 1st, 2004 as the Chair of the Department of Electrical and Computer Engineering for a three-year term.

Catherine Rosenberg is on the Scientific Advisory Board of France-Telecom and was on the Board of Governors of the IEEE Communications Society from January 2007 to December 2009. She was an Associate Editor for IEEE Communications Magazine, Telecommunications Systems, and IEEE Transactions on Mobile Computing, and served as IEEE Communications Surveys and Series co-Editor for the Series on Adhoc and Sensor Networks for IEEE Communications Magazine.