Cluster-Based Resource Management in OFDMA Femtocell Networks With QoS Guarantees

Abbas Hatoum, Rami Langar, *Member, IEEE*, Nadjib Aitsaadi, *Member, IEEE*, Raouf Boutaba, *Fellow, IEEE*, and Guy Pujolle, *Senior Member, IEEE*

***Abstract*—Recently, operators have resorted to femtocell networks to enhance indoor coverage and increase system capacity. Nevertheless, to successfully deploy such solution, efficient resource-allocation algorithms and interference mitigation techniques should be deployed. The new applications delivered by operators require large amounts of network bandwidth. Although, some customers may want to pay more in exchange for a better quality of service (QoS), some others need fewer resources and can be charged accordingly. Hence, we consider an orthogonal frequency-division multiple-access (OFDMA) femtocell network serving both QoS-constrained high-priority (HP) and best-effort (BE) users. Our objective is to satisfy a maximum number of HP users while serving BE users as well as possible. This multiobjective optimization problem is NP-hard. For this aim, we propose in this paper a new resource-allocation and admission control algorithm, which is called QoS-based femtocell resource allocation (Q-FCRA), based on clustering and taking into account QoS requirements. We show through extensive network simulations that our proposal outperforms two state-of-the-art schemes [centralized-dynamic frequency planning (C-DFP) and distributed random access (DRA)] and our previous proposal, i.e., femtocell resource allocation (FCRA), in both low- and high-density networks. The results concern the throughput satisfaction rate, spectrum spatial reuse, the rate of rejected users, fairness, and computation and convergence time.**

***Index Terms*—Clustering, femtocells, orthogonal frequency-division multiple access (OFDMA), performance evaluation, quality-of-service (QoS) support, resource allocation.**

I. Introduction

FEMTOCELLS have recently appeared as a viable solution to enable broadband connectivity in mobile cellular networks. Instead of redimensioning macrocells at the base station level, the modular installation of short-range access points can grant multiple benefits, provided that interference is opportunely managed. Technically, femtocells—referred to as femto access points (FAPs)—can drastically increase download capacity, with a much higher throughput, while offloading the macrocells, under a maximum range of a few hundred meters.

Nevertheless, the design of a large femtocell network faces technical challenges mainly arising from its coexistence with the cellular network, most notably, resource allocation and interference management. The important question to answer is that, for different network densities and different sources of interference, how can available radio resources be efficiently distributed among FAPs and the macrocell, while satisfying desired performance criteria.

There are typically two types of resource-allocation schemes that account for macrocell and femtocell coexistence: shared-spectrum [2]–[7] and split-spectrum [8]–[13] schemes. In the first case, femtocells use the same frequency band as macrocells. This results in a more dynamic resource allocation, but the interference from macrocells may seriously degrade the performance. Indeed, with a shared spectrum, femtocells lose the original advantages of resource reuse, as demonstrated in [14]. In addition, coordination mechanisms between FAPs and macrocells are needed to manage cross-layer interference. Such mechanisms may add scalability and security issues and may be counterproductive whenever there is limited availability of backhaul bandwidth [11].

On the other hand, in the case of split-spectrum schemes, FAPs use different frequency bands than those employed by macrocells, which can drastically simplify the interference management and resource allocation. It has been shown in [11] that, under an adaptive split-spectrum approach, it is possible to reach optimal area spectral efficiency. Although there are methods proposed to alleviate the macro–femto interference, interference mitigation between femtocells has not drawn much attention and, thus, forms the focus of this work. In particular, congestion cases in which femtocell demands exceed the available bandwidth pose an important challenge.

There are mainly three strategies for femtocell access, namely, closed-, open-, and hybrid-access strategies [15]. A femtocell in a closed-access mode can only be accessed by authorized femtocell users. The restriction is considered because the performance of a subscribed femtocell user can be degraded if the resources of the femtocell are shared with other nonsubscribed users in the environments where the capacity of wireless links or backhaul is limited. On the other hand, the open-access mode allows arbitrary nearby users to access the

Manuscript received September 21, 2012; revised February 18, 2013, May 24, 2013, and September 18, 2013; accepted October 8, 2013. Date of publication November 8, 2013; date of current version June 12, 2014. This paper was presented in part at the IEEE International Conference on Communications, Ottawa, ON, Canada, June 10–15, 2012 [1]. The review of this paper was coordinated by Prof. Y. Qian.

A. Hatoum, R. Langar, and G. Pujolle are with the Laboratoire d'informatique de Paris 6 (LIP6), University of Pierre and Marie Curie, Paris 75005, France (e-mail: abbas.hatoum@lip6.fr; rami.langar@lip6.fr; guy.pujolle@lip6.fr).

N. Aitsaadi is with the Laboratory of Image, Signal, and Intelligent Systems, University of Paris-Est Créteil Val de Marne, 94010 Créteil, France (e-mail: nadjib.aitsaadi@u-pec.fr).

R. Boutaba is with the School of Computer Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: rboutaba@uwaterloo.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/TVT.2013.2290125

FAPs with no restrictions. In a hybrid-access mode, a limited amount of the FAP resources are available to all users, whereas the rest are operated in a closed-subscriber-group manner. Some previous works [16], [17] have shown that the deployment of open- or hybrid-access femtocells can improve the system-wide performance. Hence, in this paper, we consider such access methods.

In this context, we propose in this paper a new resource-allocation and admission control algorithm, which is called *quality of service (QoS)-based femtocell resource allocation* (Q-FCRA), taking into account user QoS requirements. Q-FCRA relies on our previous work [18], where we proposed a cluster-based resource-allocation algorithm for orthogonal frequency-division multiple-access (OFDMA) femtocell networks, without taking into account QoS differentiation between users. In this paper, we extend our previous proposal [18] to support QoS. Indeed, we propose to distinguish between home femto-users and roaming femto-visitors in the resource-allocation process. The former users, which are referred to as high-priority (HP) users, have a strictly higher priority than the latter users, which are referred to as best-effort (BE) users. Our objective is thus to associate the best spectrum set of frequency/time resources with each FAP to maximize the number of satisfied HP users and serve as better as possible the BE users. To achieve this, we formulate the joint resource-allocation and admission control problem as a multiobjective optimization problem. The first objective is to maximize the set of admitted HP users to guarantee the feasibility of the allocation problem. The second objective is to allocate as better as possible the remaining resources to BE users.

To gauge the effectiveness of our proposal, we compare the Q-FCRA algorithm with two prominent existing strategies: centralized-dynamic frequency planning (C-DFP [8]) and distributed random access (DRA [10]), as well as our previous work (FCRA [18]). Evaluation and comparison metrics include the satisfaction rate of the required throughput, spectrum spatial reuse (SSR), rate of rejected users, fairness, and computation time. We first study the benefit of Q-FCRA when the users are static using several arbitrary FAP topologies and under various interference scenarios. Then, we incorporate user mobility to show its impact on the proposed scheme. For this purpose, we propose a new indoor mobility model, which is formulated as a three-state Markov chain, namely, "Static," "Move," and "Off" states, between which a user alternates. To the best of our knowledge, we are the first to introduce such mobility model in the femtocell performance analysis. The obtained results show that Q-FCRA outperforms existing approaches in both static and mobile environments and under various network densities and interference scenarios.

The remainder of this paper is organized as follows: Section II presents an overview of related works. In Section III, we present the network and mobility models and formulate our joint resource-allocation and admission control problem. Section IV introduces the Q-FCRA algorithm, followed by a description of the evaluation metrics in Section V. Simulation results are presented in Section VI. Finally, Section VII concludes this paper.

II. Related Work

Resource allocation in OFDMA femtocell networks has recently received significant attention. The general objective pursued is the computation of efficient allocation of time–frequency resource blocks, while accounting for cross-layer interference (interference between the macrocell and FAPs) and colayer interference (interference between FAPs).

As stated earlier, two main directions are evidenced: shared-spectrum and split-spectrum schemes. In the first, cross-layer interference needs to be managed; a proposal in this direction is [6], where Lien *et al.* present a cognitive approach for interference mitigation between the macrocell and FAPs and a game-theoretic strategy to distribute the remaining resources among FAPs. They also consider QoS guarantees in terms of delay constraints for user's applications. However, cognitive radio still present design issues in the implementation of the required functions, since additional hardware support in both macrocells and FAPs is required. In the second direction, an orthogonal channel assignment eliminates cross-layer interference by dividing the spectrum into two independent fragments. A number of related proposals have been made in the literature.

In [8], Lopez-Perez *et al.* outline the requirements of two centralized approaches, namely, orthogonal assignment algorithm and C-DFP. In the first approach, the spectrum is divided into two independent sets S_M and S_F used by the macrocells and femtocells, respectively, to maximize the satisfaction of the required QoS. However, this scheme does not take into account the femto-to-femto interference, which remains an important issue for indoor performance, particularly when femtocells are densely deployed. For C-DFP, a subchannel broker receives demands and interference information from the femtocells and/or the macrocells, to compute the best resource allocation; the tradeoff is between optimality and computational complexity. This scheme can easily converge to the optimum. However, it is practicable only for small-sized femtocell networks.

A fractional frequency reuse technique that adjusts the frequency reuse factor to alleviate interfemto interference is presented in [9]. In this case, femtocells are grouped depending on the amount of reciprocal interference by a centralized femtocell gateway that determines the minimum number of orthogonal subchannels for each group and adjusts the transmit power of each femtocell based on the received signal strength. In [19], the fractional reuse approach is analyzed. Two zones, i.e., an interior and an exterior, and a number of reserved channels for each femtocell are then required. It is shown that the computation of the zone boundary depends on an effective signal-to-interference-plus-noise ratio (SINR) calculation.

Sundaresan and Rangarajan in [10] propose a distributed resource-allocation algorithm, namely, DRA, which is more appropriate for medium-size networks. The resources are split between macrocells and femtocells based on the gradient ascent/descent heuristic. Once the resource set dedicated to femtocells is determined, each femtocell locally runs DRA to reserve a set of resources using a randomized hashing function. To do so, each femtocell divides the resources into blocks proportional to the number of interfering neighbors. It is shown

that this algorithm is fully distributed with an acceptable worst case performance guarantee.

Similarly, Chandrasekhar and Andrews in [11] propose a decentralized Frequency-ALOHA, F-ALOHA spectrum allocation strategy for two-tier cellular networks. The proposal is based on a dynamic partition of the spectrum between the macrocell and femtocells. Once computed, each femtocell accesses a random subset of the candidate frequency subchannels. The probability to reserve a subchannel depends on the set of its interfering femtocells. It is shown that this approach optimizes the area spectral efficiency. However, one should note that due to their pseudo-random nature, these two approaches (i.e., [10] and [11]) cannot guarantee QoS in a realistic scenario.

In [12], Garcia *et al.* propose a fully distributed and scalable algorithm for interference management in LTE-Advanced environments, namely, Autonomous Component Carrier Selection (ACCS). In this method, each FAP always has one active component carrier, denoted by the primary component carrier, which is selected according to the computed path loss. If the offered QoS in terms of bandwidth is not sufficient, a FAP tries to reserve more carrier components (i.e., secondary component carrier) without deteriorating the QoS of neighboring FAPs. It has been shown that ACCS improves the experience of all users without compromising the overall cell capacity. However, the scheme is highly correlated with environmental sensing since it mainly relies on measurement reports. In addition, ACCS does not allocate time–frequency slots but only subcarriers, which can be expensive and penalizing in terms of bandwidth.

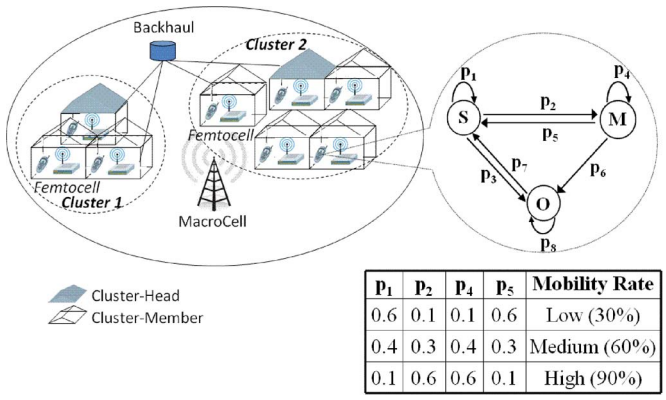
Arslan *et al.* in [13] present a femtocell resource management system for interference mitigation in OFDMA networks, namely, FERMI. It is composed of two functional modules. The first module uses coarse-level measurements to classify clients into two categories: those that need resource isolation (i.e., need orthogonal subchannels) and those that do not. The second module assigns OFDMA subchannels to the different femtocells in a near-optimal fashion.

So far, the aforementioned schemes have not taken into account QoS differentiation between users. However, some studies in standard cellular systems and, recently, in femtocell networks have considered QoS-aware resource-allocation algorithms. A selection of relevant works is discussed in the following.

Ergen *et al.* in [20] consider the problem of assigning a set of subcarriers and determining the number of bits to be transmitted for each subcarrier. They introduce an iterative multiuser bit and power allocation scheme to meet the QoS requirements. Their objective is to minimize the total transmit power by allocating subcarriers to the users and then determine the number of bits transmitted on each subcarrier.

In [21], Choi *et al.* develop a QoS-aware selective feedback model where each user chooses those channel sets that meet its QoS requirements by exploiting user diversity. Given the feedback channel sets for each user, the base station distributes channels to each user. Their objective is to maximize the number of users or the sum of users' utility values.

Both subcarrier and power-allocation methods are also presented in [22], where users are differentiated by service type to fulfill the QoS requirements. In this case, the aim is to minimize



p_1	p_2	p_4	p_5	Mobility Rate
0.6	0.1	0.1	0.6	Low (30%)
0.4	0.3	0.4	0.3	Medium (60%)
0.1	0.6	0.6	0.1	High (90%)

Fig. 1. Network and mobility models.

the power sum required to satisfy all HP users so that the power dedicated to BE users is maximized. However, such scheme needs network coordination between base stations to improve users' performance at cell borders.

Recently, Liang *et al.* in [23] have proposed a greedy algorithm for physical resource block allocation with QoS classes for different services. The proposed approach assumes the use of a central controller and gateway, i.e., the so-called FMS, coordinating network access for femto installations and responsible for resource allocation. However, as in [8], it is practicable only for small-sized femtocell networks. In addition, for 3GPP LTE specification [24], i.e., the Long-Term Evolution femto radio nodes (e)HomeNodeB, the gateway network element is optional with respect to the connection to the management platform. This fact strengthens the need for self-organizing management and optimization technology, as shown in our study.

To this end, we propose in this paper a joint resource-allocation and admission control strategy for femtocell networks, which provides QoS guarantees for HP users and maximizes the throughput for BE users. We formulate the problem as a multiobjective optimization problem, and to solve it, we use the concept of irreducible infeasible set (IIS) minimizing the sum of elastic variables for HP users and a Min–Max optimization problem for BE users (see Section IV-B). It is worth noting that, while the adaptive split-spectrum approach [11] is used to determine the best allocated resources dedicated to femtocells, our approach is used to optimize the resource distribution among femtocells. Both approaches are complementary and needed to achieve a fair resource allocation across the entire network with high spatial reuse.

III. System Model

Here, we first define the network and mobility models used in our analysis; then, we formulate our joint resource allocation and admission control as a multiobjective optimization problem.

A. Network Model

We consider an OFDMA (e.g., LTE) femtocell's network consisting of several FAPs representing residential or enterprise networks, as shown in Fig. 1. As in [8], [10], [11], [13], and [23], we adopt an orthogonal channel assignment

that eliminates the cross-layer interference between femtocells and the macrocell. In our study, we focus on the downlink communications based on OFDMA, whose frame structure can be viewed as time–frequency resource blocks, which are also called tiles.

A tile is the smallest unit of resource that can be assigned to a user and corresponds to 0.5-ms and 180-KHz frequency band. According to the LTE specification [24], scheduling is done on a subframe basis for both the downlink and the uplink. Each subframe consists of two equally sized slots of 0.5 ms in length. A certain number of users attach to each FAP; user demands represent the required bandwidth, expressed in number of required tiles. The relation between required tiles of user u (D_u) and the throughput requirement (TP_u^{req}) can be written as follows:

$$D_u = \left\lceil \frac{TP_u^{\mathrm{req}}}{\psi \cdot eff_u} \right\rceil \tag{1}$$

where $\psi = (SC_{\mathrm{ofdm}} \cdot SY_{\mathrm{ofdm}})/T_{\mathrm{subframe}}$ is a fixed parameter that depends on the network configuration; SC_{ofdm} and SY_{ofdm} are the numbers of subcarriers and symbols per tile, respectively; and T_{subframe} is the frame duration in time units. In the LTE specification [24], $SC_{\mathrm{ofdm}} = 12$, $SY_{\mathrm{ofdm}} = 7$, and $T_{\mathrm{subframe}} = 0.5$ ms. Parameter eff_u is the efficiency (bits/symbol) of the used modulation and coding scheme (MCS).

In this paper, for simplicity, we consider a fixed transmission power for all FAPs, as in [8], [10], [11], [13], and [23]. In addition, we do not consider adaptive MCS. Indeed, our aim is not to show the effectiveness of using link adaptation in our cluster-based approach but rather to study how the tiles are effectively assigned to femto-users, taking into account QoS requirements and interference levels. Note that using link adaptation will further enhance the performance of our approach. This will be explored in a future work.

In our study, two types of users are considered: HP users who require a certain QoS guarantee in terms of bandwidth and BE users. Note that the differentiation between users is not based on the different QoS demands. Instead, users are differentiated based on parameters such as the plan they subscribe with the operator; or if they are the owner of the FAP (i.e., home femto-users) compared with a visitor connecting to a FAP (i.e., roaming femto-visitors) in open or hybrid access. In this case, home femto-users have a strictly higher priority than roaming femto-visitors.

As previously mentioned, in urban dense environment, we expect that, often, the sum of demands of the FAPs exceeds the available resources. Therefore, our objective is to find, for such contention situations, an effective resource allocation that takes care of throughput expectations while controlling the interference between femtocells and, at the same time, taking into account the QoS requirements of both HP and BE users. It is worth noting that HP users of the network have fixed QoS requirements. Since the total amount of network resources is limited, an admission control strategy for HP users is then needed. An HP user is indeed admissible only when the network has sufficient resources to meet the QoS requirements.

It is worth noting that this QoS differentiation between users can be adopted even in close subscriber group (CSG) access method. Indeed, according to the 3GPP technical report [25], to mitigate FAP interference and to prevent free FAP usage by neighbors, one option is to change the CSG ID dynamically between its default CSG ID (which is assigned when it is deployed) and a dedicated CSG ID (which is configured by the operator). In this case and as reported in [25], the closed-access FAP becomes accessible to each passing-by user, for a period of time, thus alleviating its interference with the macrocell. However, one should note that, in this case, a QoS-aware resource allocation is required to differentiate between home femto-users and passing-by femto-visitors.

B. Mobility Model

We propose a mobility model to represent indoor user movement. The proposed model consists of three states/phases, namely, "Static" (S), "Mobile" (M), and "Off" (O) states, between which a user alternates. State "O" means that a communication session is terminated. In states "S" and "M," a user is able to send/receive data to/from its associated femtocell. Our indoor mobility model can be represented by the simple three-state Markov chain shown in Fig. 1, where the movement performed in state "M" follows a random walk [26]. Transition probabilities control the users' mobility rate, the session duration (SD), and the network load. Indeed, p_3 and p_6 control the SD, and p_7 determines the network load. On the other hand, by varying probabilities p_1, p_2, p_4, and p_5, several mobility rates can be achieved, as reported in Fig. 1. These parameters can be used by Q-FCRA in the determination, for instance, of the intracluster resource-allocation computation epoch (see Section IV-B).

C. Problem Formulation

Let $\mathcal{F}$ be the set of FAPs, $\mathcal{H}$ the set of HP users, and $\mathcal{B}$ the set of BE users in the network. Moreover, let $\mathcal{I}_f$ and $\mathcal{I}_u$ be the set of interfering femtocells of the FAP F and the interfering set of user u, respectively. More precisely, $\mathcal{I}_f$ corresponds to the set of femtocells composed of F and the femtocells causing interference to F. This set is determined using the minimum required SINR values and the indoor path loss model. Indeed, each user within the FAP F boundary calculates the ratio of the received signal from F to the signals received from all surrounding/neighboring FAPs. If this ratio is lower than the minimum required SINR, the user notifies its serving FAP. The corresponding neighboring FAPs will be then considered as interferers for F and will belong to $\mathcal{I}_f$. Note that user v interferes with user u and should not be assigned the same resources (i.e., tiles) if u and v are attached to the same FAP or user v belongs to a different FAP that creates interference on user u. Hence, the set $\mathcal{I}_u$ can be written as follows. For each user u attached to the FAP F, $\mathcal{I}_u = \{v \neq u | v \in F \cup \mathcal{I}_f\}$. It is worth noting that users use the received SINR to calculate the channel quality indicator they report to the network. Based on these measurement reports, a FAP can decide to reallocate the tile or, in the case of using link adaptation, allocate a different MCS.

In addition, we denote by D^u_{HP} and D^v_{BE} the traffic demand of HP user $u \in \mathcal{H}$ and BE user $v \in \mathcal{B}$, respectively. We also define for each HP user u (respectively, BE user v) the binary resource-allocation vector, denoted by Δ^u_{HP} (respectively, Δ^v_{BE}), with 1 or 0 in position j according to whether tile j is used or not.

As stated earlier, our objective is to find the optimal resource allocation of a set of tiles in each FAP to deliver users' data, while minimizing the interference between FAPs and, at the same time, providing QoS guarantees for HP users and maximizing the throughput for BE users. Due to the limited network capacity, if the QoS requirements of HP users exceed the available resources, then satisfying all HP users becomes infeasible. Hence, we first define the set of admitted HP users in the network, which is denoted by $\mathcal{H}^* \subseteq \mathcal{H}$, for which the QoS requirements are fully satisfied. Our first objective will be then to maximize the cardinality of set $\mathcal{H}^*$.

In addition, for each BE user v, we define a variable $G_{\mathrm{BE}}(v)$, which represents the gap between the required and the allocated resources of v. That is, $G_{\mathrm{BE}}(v) = (D^v_{\mathrm{BE}} - \sum_{j=1}^{M} \Delta^v_{\mathrm{BE}}(j)/D^v_{\mathrm{BE}})$, where M denotes the number of available resources (i.e., tiles) in the network. Our second objective will be then to minimize the maximum value of G_{BE}. As such, we will guarantee a certain degree of fairness while serving the maximum number of BE users, as will be shown in the simulation results in Section VI. Given the set of interfering users I_u, our joint resource-allocation and admission control problem for HP and BE users can be formulated as shown in Problem 1.

Problem 1 Joint resource allocation and admission control for HP and BE users

$$\max |\mathcal{H}^*|$$
$$\min[\max_{v\in\mathcal{B}} G_{\mathrm{BE}}(v)]$$
subject to:
(a) $\forall u \in \mathcal{H}^*$: $\sum_{j=1}^{M} \Delta^u_{\mathrm{HP}}(j) = D^u_{\mathrm{HP}}$
(b) $\forall v \in \mathcal{B}$: $\sum_{j=1}^{M} \Delta^v_{\mathrm{BE}}(j) \le D^v_{\mathrm{BE}}$
(c) $\forall j = 1, \ldots, M$,
$\forall u \in \mathcal{H}^* \cup \mathcal{B}, \forall v \in \mathcal{I}_u$, : $\Delta^u_{\cdot}(j) + \Delta^v_{\cdot}(j) \le 1$
(d) $\forall j, \quad \forall u \in \mathcal{H}^* \cup \mathcal{B}$: $\Delta^u_{\cdot}(j) \in \{0, 1\}$

In this problem, condition (a) denotes that the resource scheduler must guarantee that admitted HP users are fully satisfied. Condition (b) denotes that BE users cannot obtain more than the required data rate. Inequality (c) ensures that two interfering users cannot use the same tiles. Condition (d) indicates that $\Delta_{\cdot}(j)$ is a binary variable.

Problem 1 is a multiobjective optimization problem and has been proved to be NP-hard [27]. To solve it, we propose first to subdivide it into subproblems by means of clustering. The corresponding problem will then be sequentially solved. That is, we will try to satisfy HP users first and then resolve for BE users with the remaining resources. This approach drastically reduces the time complexity of the resource-allocation problem and implies successive provisioning steps, as described in Section IV.

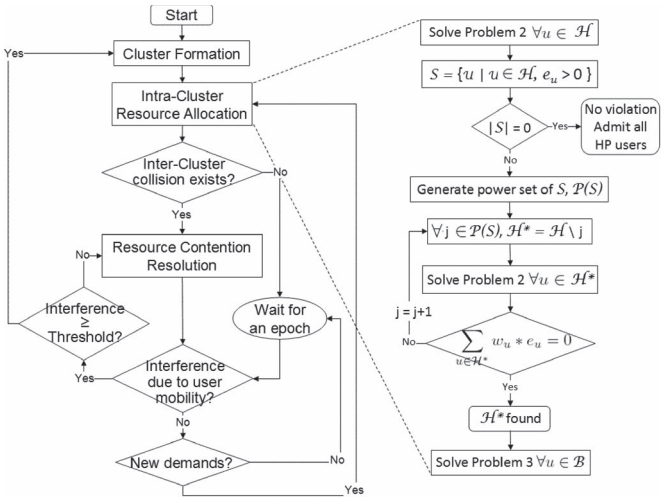


Fig. 2. Flowchart of Q-FCRA.

IV. Proposal: QoS-Based Femtocell Resource Allocation Algorithm

Here, we present our Q-FCRA algorithm for OFDMA femtocell networks. Similar to [18], Q-FCRA consists of three main stages: 1) cluster formation; 2) intracluster resource allocation; and 3) intercluster resource contention resolution (see flowchart in Fig. 2). In what follows, we present these three stages focusing, in particular, on the second stage.

A. Cluster Formation

When powered on, a FAP will listen to surrounding transmissions (i.e., neighboring FAPs' control channel and reference signal transmissions) and gather information through measurements collected from users attached to it or via a receiver function within the FAP, which is also called "Sniffer" [25]. Based on this information, the FAP F can compute the number of interfering femtocells (i.e., $|\mathcal{I}_f|$, called interference degree) and transmit it along with its physical cell identity to each one of them.[1] Therefore, each FAP will have a list containing the interference degree of neighboring femtocells and will decide whether it is a cluster head (CH) or is attached to a neighboring cluster. The CH election algorithm can be described as follows.

- Each femtocell elects the CH as the one with the highest interference degree among its one-hop neighbors.
- If it is not a CH itself, the femtocell acts as a cluster member (CM) of a CH chosen by its immediate neighbors.
- If more than one unique CH is chosen by the neighboring femtocells, the one with the highest interference degree is elected as the CH to minimize the collision of tiles between femtocells (if equal degrees, a random tie-break is used).
- If no CH is chosen by the neighboring femtocells (i.e., all neighbors act as CMs and are already associated to other clusters), the femtocell is attached to the cluster of the neighbor with the highest interference degree. It is

[1]The information exchange can be done using "over-the-air via UE" approach, as detailed in the 3GPP technical report [25].

worth noting that to avoid large cluster size due to the attachment of such femtocells to neighboring clusters, we set a threshold on its size denoted by CS_{th}. When the threshold is reached, the corresponding femtocell will act as an isolated CH.

The cluster formation stage is more formally described by the pseudocode in Algorithm 1. It is worth noting that the formation of a given cluster is renewed only when the interference caused by users' mobility exceeds a given threshold, as shown in the flowchart in Fig. 2 (in our simulations, this threshold is set to 5 dB). In addition, we note that time synchronization between femtocells is necessary to enable accurate resource-allocation decisions. The most and simple used one is timing from the Internet. Indeed, once turned on, and before initiating any communication, femtocells get synchronized to the cellular core network using an asymmetric communication link, such as DSL technologies, due to an enhanced version of IEEE 1588 [28].

Algorithm 1 Cluster Formation Algorithm for a given Cluster Size threshold (CS_{th}) - Femtocell $\mathcal{F}_a$

$\mathcal{F}_a$ establishes a list of interfering neighbor femtocells;
$\mathcal{F}_a$ sends the list to its one-hop neighbors through a backhaul link;
if $\mathcal{F}_a$ has the highest degree of interfering neighbors **then**
  $\mathcal{F}_a$ elects itself as a CH and informs its one-hop neighbors;
  Set property IsClusterHead = True;
  Set Cluster Size $CS_a = 1$;
else
  if $\mathcal{F}_a$ has neighboring CHs **then**
    Sort the list of neighboring CHs decreasingly according to their interference degree;
    $\mathcal{F}_a$ selects the first from the list (i.e., the highest interfering neighbor CH), $\mathcal{F}_b$;
    if $CS_b \leq CS_{\text{th}}$ **then**
      $\mathcal{F}_a$ attaches to the cluster administrated by $\mathcal{F}_b$;
      Set property HasClusterHead = True;
      $CS_b = CS_b + 1$;
    else
      Remove $\mathcal{F}_b$ from the list;
      if the list $\neq$ Null **then**
        Go to step 10;
      else
        Go to step 30;
      end if
    end if
else
    $\mathcal{F}_a$ selects the highest interfering neighbor femtocell and sends attachment request to its corresponding CH, $\mathcal{F}_c$;
    if $CS_c \leq CS_{\text{th}}$ **then**
      $\mathcal{F}_a$ attaches to the cluster administrated by $\mathcal{F}_c$;
      Set property HasClusterHead = True;
      $CS_c = CS_c + 1$;
    else
      $\mathcal{F}_a$ acts as an isolated CH;
    end if
  end if
end if

B. Intracluster Resource Allocation

Once the femtocell network is partitioned into clusters, the second step is to jointly allocate resources to all FAPs within each cluster, taking into account QoS requirements of attached users. To achieve this, each CM reports to its corresponding CH the required resources to satisfy its user's demands. Then, each CH tries to individually resolve the original problem (Problem 1) for every epoch δ_t, since it depends on the arrival/departure process of end users residing in the cluster. Note that the envisioned order of an epoch could be minutes, hours, or days, depending on whether the mobility of femtocell users within a cluster is globally high, medium, or low. δ_t can be expressed as follows:

$$\delta_t = \Pi_S \times 1/\mu \tag{2}$$

where Π_S denotes the steady-state probability for state "S" of the Markov chain shown in Fig. 1, and $1/\mu$ is the mean sojourn time of a femto user within its corresponding FAP.

A good solution of the original problem could be attained by sequentially resolving the two objectives of Problem 1. In the following, we present our approach to resolve the resource-allocation problem for HP users first and then for BE users. It is worth noting that, since the obtained clusters' size is not large (in our simulations, $CS_{\text{th}} = 10$), the CH resolution using a solver such as "IBM ilog cplex" [29] would still converge within a short time period (as will be shown in the simulation results in Section VI). This allows femtocells to serve their attached users in a timely manner.

1) HP Users Admission Control and Resource Allocation: As stated earlier, we need to choose a subset of HP users for which the allocation problem is feasible. Since the objective is to maximize the number of satisfied HP users, the cardinality of such a subset has to be the maximum of all such subsets. This problem is equivalent to the IIS problem. An IIS is an infeasible set of constraints of which any proper subset is feasible. That is, if we remove any one constraint from an IIS, the IIS will be feasible. In linear programming, this set is often difficult to determine. Hence, a useful approach called "elastic programming" was introduced by Brown and Graves [30]. It consists of adding an extra "slack" variable allowing constraints to "relax" to increase the feasibility region. In other words, if an HP user cannot fulfill its demands with the available resources, it will use a certain elastic variable to complete its requirements. Thus, for each HP user u, we introduce an elastic variable e_u. To locate the inconsistent constraints, Chinneck and Dravnieks [31] proposed to create a new objective function—minimize the sum of elastic variables—and then perform "filtering," where the constraints having elastic variables greater than zero form the set of inconsistent constraints. Hence, the optimization problem for HP users will be formulated as follows:

Problem 2 Minimize the sum of elastic variables for HP users

$$\begin{aligned}&\min \textstyle\sum_{u\in\mathcal{H}} w_u \times e_u\\ &\text{subject to:}\\ &\text{(a) } \forall u \in \mathcal{H}: \quad \textstyle\sum_{j=1}^{M} \Delta_{\mathrm{HP}}^{u}(j) + e_u \geq D_{\mathrm{HP}}^{u}\\ &\text{(b) } \forall j,\ \forall u \in \mathcal{H},\ \forall v \in \mathcal{I}_u: \quad \Delta_{\mathrm{HP}}^{u}(j) + \Delta_{\mathrm{HP}}^{v}(j) \leq 1\\ &\text{(c) } \forall u \in \mathcal{H}: \quad e_u \geq 0\end{aligned}$$

In Problem 2, $\mathcal{H}$ represents the set of HP users within the cluster, and $w_u \in \Re_+$ are weighting coefficients used to set priority levels between different HP users. When $w_u = 1,\ \forall u$, then all users are given equal priority.

Note that an optimal value of an elastic variable e_u^* should be zero for the corresponding HP user to be fully satisfied. On the other hand, a nonzero solution indicates the need for more resources than available in the network to satisfy the corresponding HP user. Therefore, we will admit into the network only those HP users whose corresponding elastic variables e_u reach zero. The complete resolution algorithm for HP users is described as follows and is represented in the flowchart in Fig. 2.

- First, each CH resolves Problem 2 for all HP users within the cluster.
- Then, we determine set $\mathcal{S}$ of HP users for which the elastic variables are greater than zero.
- The power set of $\mathcal{S}$, which is denoted by $\mathcal{P}$(S), is generated. It is composed of all subsets of $\mathcal{S}$. That is, assuming $\mathcal{S}$ is a finite set with cardinality $|\mathcal{S}| = n$, then $\mathcal{P}$(S) is finite, and its cardinality $|\mathcal{P}(S)| = 2^n$. For example, if $\mathcal{S} = \{1, 4, 6\}$ representing HP users with nonzero elastic variables, then $\mathcal{P}(S) = \{\{1\}, \{4\}, \{6\}, \{1, 4\}, \{1, 6\}, \{4, 6\}, \{1, 4, 6\}\}$.
- Afterward, Q-FCRA removes the elements $\mathcal{X}$ within $\mathcal{P}$(S) one at a time starting with the lowest cardinality subsets and resolves Problem 2 for the remaining set $\mathcal{H}^*$ of HP users, where $\mathcal{H}^* = \mathcal{H} - \mathcal{X}$.
- If $\sum_{u\in\mathcal{H}^*} w_u \times e_u > 0$, the previously removed element is reinserted, since its removal did not allow the feasibility of the problem, and the next element within $\mathcal{P}$(S) is removed. Problem 2 is resolved again on the new set $\mathcal{H}^*$.
- The process is stopped if $\sum_{u\in\mathcal{H}^*} w_u \times e_u = 0$.

At the end of this process, set $\mathcal{H}^*$ of admitted HP users (for which the original problem is feasible) and the corresponding allocation matrix $\mathbf{A}_{\mathrm{HP}}$ of dimensions $|\mathcal{H}^*| \times M$ are determined. The next step is now to allocate the remaining resources to BE users.

2) BE Users Resource Allocation: The set of tiles that BE users can have access to depends on the allocation of interfering HP users within the cluster. Thus, we denote by $\mathbf{I}_{\mathrm{BE,HP}}$ the interference matrix of dimensions $|\mathcal{B}| \times |\mathcal{H}^*|$ between BE and admitted HP users, with 1 or 0 in position (m, n) according to whether BE user m interferes with HP user n or not. (Note that $\mathcal{B}$ represents in this case the set of BE users within the cluster.)

The resulting matrix $\mathbf{R}_{\mathrm{BE}} = \mathbf{I}_{\mathrm{BE,HP}} \times \mathbf{A}_{\mathrm{HP}}$ of dimensions $|\mathcal{B}| \times M$ can be calculated such that in position (i, j), $r_{ij} = \sum_{k=1}^{|\mathcal{H}^*|} \mathbf{I}_{\mathrm{BE,HP}}(i, k) \times \mathbf{A}_{\mathrm{HP}}(k, j)$. Hence, its complementary matrix $\overline{\mathbf{R}}_{\mathrm{BE}}$ can be defined as

$$\overline{r}_{ij} = \begin{cases} 1, & \text{if } r_{ij} = 0\\ 0, & \text{if } r_{ij} \geq 1.\end{cases}$$

Note that $\overline{r}_{ij} = 1$ indicates that tile j can be allocated to BE user i. Hence, the BE users resource-allocation problem can be formulated as follows:

Problem 3 Resource allocation for BE users

$$\begin{aligned}&\min[\max_{i\in\mathcal{B}} G_{\mathrm{BE}}(i)]\\ &\text{subject to:}\\ &\text{(a) } \forall\, i \in \mathcal{B}: \quad \textstyle\sum_{j=1}^{M} \Delta_{\mathrm{BE}}^{i}(j) \leq D_{\mathrm{BE}}^{i}\\ &\text{(b)} \forall\, j = 1, \ldots, M,\\ &\forall\, i \in \mathcal{B},\ \forall\, k \in \mathcal{I}_i: \quad \Delta_{\mathrm{BE}}^{i}(j) + \Delta_{\mathrm{BE}}^{k}(j) \leq 1\\ &\text{(c) } \forall\, j = 1, \ldots, M,\ \forall\, i \in \mathcal{B}: \quad \Delta_{\mathrm{BE}}^{i}(j) \leq \overline{r}_{ij}\end{aligned}$$

where condition (c) ensures that a BE user cannot use the same tile as an interfering HP user.

C. Intercluster Resource Contention Resolution

According to the previous stage, users at the edge of two neighboring clusters might still interfere when they operate on the same resources. This could indeed happen since each CH resolves the resource-allocation problem independently from its neighboring clusters. Consequently, two interfering femtocells attached to different clusters could use the same allocated tile. To resolve such collisions, a simple yet efficient mechanism can be realized and described as follows.

- Each user suffering from contention will send a feedback report (as of 3GPP specifications [25]) to its associated femtocell to notify it about the collision on the selected tile.
- Each femtocell tries to resolve contention on the collided tiles by sampling a Bernoulli distribution. Accordingly, it decides whether the attached user would keep using the tile or would remove it from the allocated resources.

It is worth noting that if collision occurs, Q-FCRA converges to a stationary allocation within a short time period, as will be shown in Section VI. This makes our solution practically feasible.

V. Performance Metrics

The performance of our proposal is evaluated considering the following QoS metrics: rate of rejected users, throughput satisfaction rate (TSR), SSR, fairness, and computation time.

A. Rate of Rejected Users

This metric represents the percentage of HP and BE users not admitted into the network. Recall that, once accepted, HP users are completely satisfied, whereas for BE users, their satisfaction degree will be maximized.

B. TSR

TSR denotes the degree of satisfaction of a user with respect to the requested resources. For each user u attached to a FAP $\mathcal{F}_a \in \mathcal{F}$, $TSR(u)$ is defined as the ratio of the allocated number of tiles to the requested tiles and can be expressed as follows:

$$\forall\, u, \qquad TSR(u) = \left(\sum_{j=1}^{M} \Delta_{\cdot}^{u}(j)\right) \Big/ \mathcal{D}_{\cdot}^{u}. \tag{3}$$

For a network with N users, the TSR metric can be thus given by

$$TSR = \sum_{u} TSR(u)/N. \tag{4}$$

C. SSR

SSR denotes the average portion of FAPs using the same tile within the network. Therefore, it is defined as the mean value of tiles' spatial reuse. The SSR metric can be thus expressed as follows:

$$SSR = \frac{1}{M \times |\mathcal{F}|} \sum_{k=1}^{M} \sum_{u \in \mathcal{H} \cup \mathcal{B}} \Delta_{\cdot}^{u}(k). \tag{5}$$

D. Fairness

Fairness is evaluated in terms of the fairness index [32], which determines how fairly the resources are distributed among N existing users. It is expressed as follows:

$$\beta = \left(\sum_{u=1}^{N} TSR(u)\right)^{2} \Big/ \left(N \cdot \sum_{u=1}^{N} TSR(u)^{2}\right). \tag{6}$$

E. Computation and Convergence Time

This is the time needed for the system to resolve the resource-allocation problem (using a solver) for both HP and BE users and converges to a stationary allocation (i.e., no resource contention). Recall that, in both FCRA and Q-FCRA, a Bernoulli distribution is used to resolve resource contention between users. In what follows, we give an analytical expression of the convergence time for both FCRA and Q-FCRA schemes.

Let $\mathcal{C}$ be the set of remaining collided tiles after an iteration k, and n the number of interfering femtocells that use the same tile $i \in \mathcal{C}$. Moreover, let p be the probability of success of the Bernoulli distribution.

The probability of resolving the collision on tile i can be formulated as

$$q_n = n \times p \times (1-p)^{n-1} + (1-p)^{n}. \tag{7}$$

Given the number of collided tiles $m = |\mathcal{C}|$ at the end of iteration k, the probability of convergence (denoted by P_c) at iteration $k+1$ can be recursively expressed as follows:

$$\begin{aligned} P_c(k+1, m) = &\sum_{i=1}^{m-1} \binom{m}{i} q_n^{i} (1-q_n)^{m-i} \times P_c(k, m-i) \\ &+ (1-q_n)^{m} \times P_c(k, m) \qquad \forall \quad k \geq 1 \\ P_c(1, m) = &\, q_n^{m}. \end{aligned} \tag{8}$$

Hence, the average convergence time can be given by

$$Conv\_time = \sum_{k=1}^{\infty} k \times P_c(k, m). \tag{9}$$

VI. Performance Evaluation

Here, we evaluate the efficiency of our proposal under various interference scenarios and FAP densities. C-DFP [8], DRA [10], and FCRA [18] schemes are used as benchmarks to which the Q-FCRA potential benefits are compared. The reported results are obtained using the solver "IBM ilog cplex" [29]. The number of users in each FAP and their traffic demands are varied at each simulation. We consider a typical OFDMA frame (downlink LTE frame) consisting of $M = 100$ tiles, as in [10]. This corresponds to a channel bandwidth of 10 MHz, which is most commonly used in practice (i.e., 50 tiles in the frequency domain) and one subframe of 1 ms in length (i.e., two time slots).

We consider different network sizes: 50 and 200 FAPs, representing low- and high-density networks, respectively. The FAPs are randomly distributed in a 2-D 400 m $\times$ 400 m area, with one FAP randomly placed in a 10 m $\times$ 10 m residence. Note that some of the residences might not have a FAP or the FAP might be switched off. We consider buildings with 3 $\times$ 3 residence blocks each. Buildings are separated by streets and cover the considered area. This basically represents a neighborhood. Users are uniformly distributed within the residence with a maximum number of ten users per FAP. In the case of Q-FCRA, these users are divided into four HP users with equal priority (i.e., $w_u = 1,\ \forall u \in \mathcal{H}$) and six BE users. Each user uniformly generates its traffic demand that can be directly translated to a certain number of tiles using (1), with a maximum value of 20 tiles per HP user and 10 tiles per BE user. Moreover, we consider different minimum required SINR thresholds, i.e., 10, 15, 20, and 25 dB, to show the impact of the interference level on the evaluated metrics. Based on the SINR, the path loss model given in the A1 scenario for indoor small office and residential of WINNER [33] for the frequency range 2–6 GHz, each femtocell determines the set of its interfering femtocells, depending on the received signal strength. It is worth noting that each SINR threshold corresponds to a certain MCS. For example and according to [34], 16-QAM modulation with a coding rate of 3/4 corresponds to a minimum SINR threshold between 10 and 15 dB. While higher order of MCS (e.g., 64-QAM 3/4 or 64-QAM 5/6) requires, respectively, a threshold of 17.5 and 20 dB.

In what follows, we present the corresponding simulation results for both static and mobile environment scenarios. The results are obtained over many simulation instances for each scenario, with a margin error less than 2%, and we calculate the mean value of performance metrics. We do not, however, plot the corresponding confidence intervals for the sake of presentation.

A. Static Environment Scenario

In the first scenario, we consider a static distribution of end users within the network, where users' position and traffic

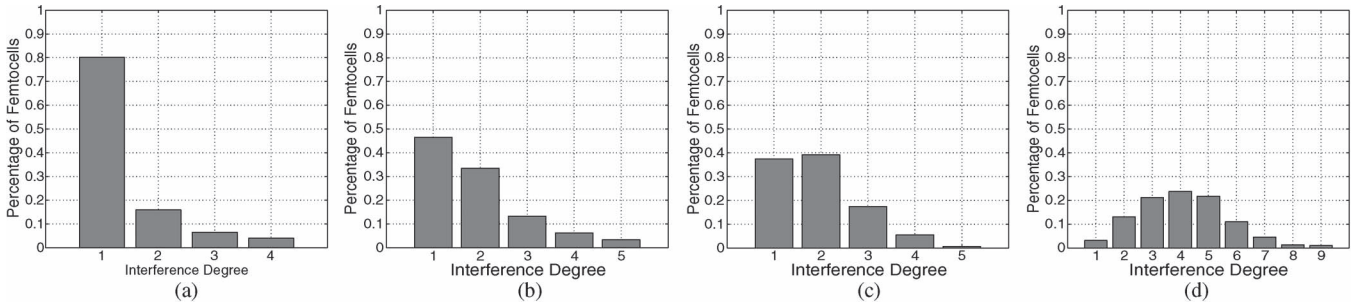


Fig. 3. Interference degree distribution for used topologies. (a) SINR = 10 dB/50 FAPs. (b) SINR = 25 dB/50 FAPs. (c) SINR = 10 dB/200 FAPs. (d) SINR = 25 dB/200 FAPs.

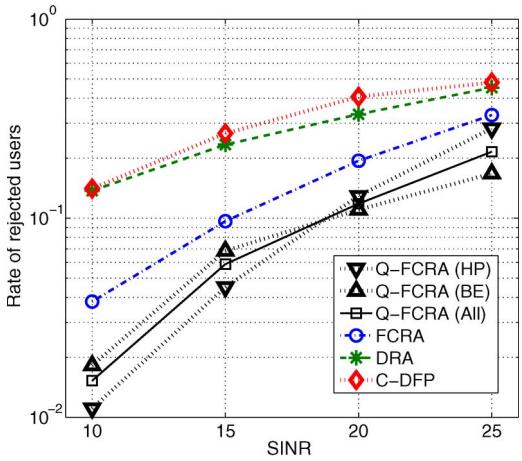


Fig. 4. Rate of rejected users.

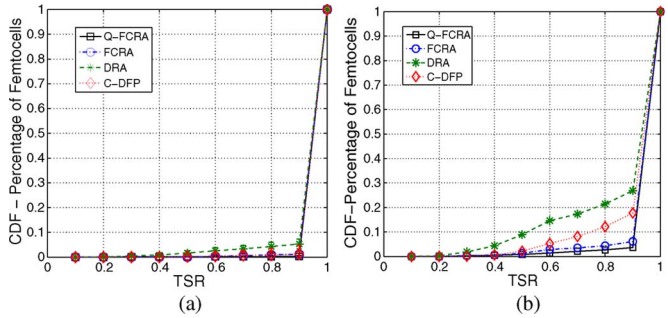


Fig. 5. CDF of TSR in low-density networks. (a) $SINR = 10$ dB. (b) $SINR = 25$ dB.

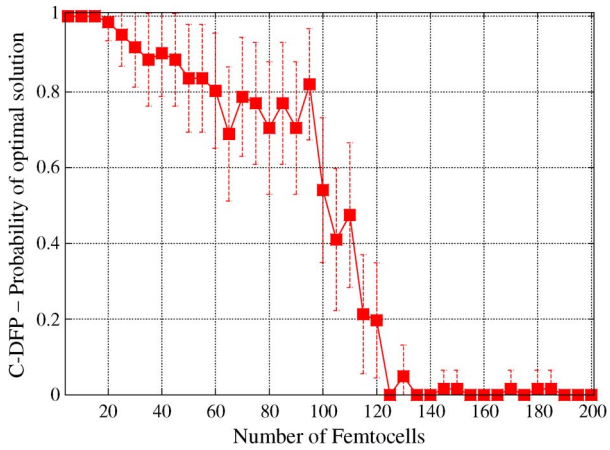


Fig. 6. Probability of finding the optimal solution for the centralized approach, i.e., $SINR = 25$ dB.

demands remain constant. Before delving into the exploration of the results, let us start by giving an idea about the topologies used in our analysis with the femtocell interference degree distribution (corresponding to the number of neighboring FAPs causing interference). As can be noticed in Fig. 3, 50-FAP topologies present a majority of isolated femtocells that do not suffer from interference, while this is no longer the case for 200-FAP topologies, which present a considerable increase in the high interference degrees with the 25-dB SINR threshold.

Let us now focus on the comparison among the different strategies based on the rate of rejected users, TSR, the SSR, and the computation and convergence time.

1) Rate of Rejected Users: Fig. 4 reports the ratio of the rejected users for the 200-FAP network case using the aforementioned allocation schemes. We detail in this figure the respective rejection ratios of both HP and BE users in the Q-FCRA scheme as well as the sum of all users in the network. In this figure, we can see that Q-FCRA allows about 98% of users to be admitted in the network (for both HP and BE users), compared with 95% for FCRA and 85% for both C-DFP and DRA in low interference levels. Whereas, with high interference, since the algorithm is strict regarding HP users' satisfaction, their rejection ratio increases; this allows, on the other hand, a higher number of BE users, with less restrictive requirements, to be served. While the rejection ratio of Q-FCRA in the worst-case scenario is still below 20%, in the two latter schemes, it reaches more than 40%.

2) TSR: Fig. 5 shows the cumulative distributed function (cdf) of the TSR for low-density networks in low and high interference levels. We can see that both Q-FCRA and FCRA converge to the optimal centralized solution (C-DFP) when the interference level is low. The reason is that in this case, the clusters constructed by our approaches often contain a small number of nodes (typically one or two FAPs). Hence, each FAP can use the entire available spectrum satisfying all the users demand. However, the performance decreases with the increase in the interference, particularly for the C-DFP method. Indeed, as shown in Fig. 6, the probability to generate the optimal solution with C-DFP, when SINR = 25 dB, is inversely proportional to the network size. Specifically, based on extensive simulations, the probability of finding the optimal solution is equal to 1 if the number of femtocells is low (i.e., $N \leq 20$). However, in

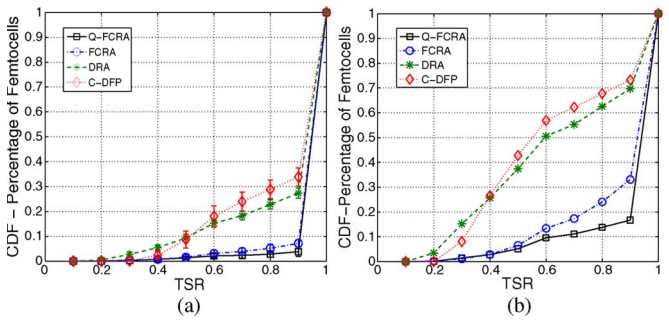


Fig. 7. CDF of TSR in high-density networks. (a) $SINR = 10$ dB. (b) $SINR = 25$ dB.

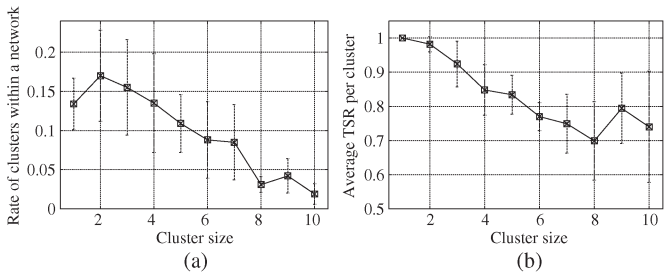


Fig. 8. Impact of cluster size on the performance of Q-FCRA in high-density networks, i.e., $SINR = 25$ dB. (a) Rate of clusters. (b) Average TSR per cluster.

a large-sized network (i.e., $N \geq 100$), this probability becomes roughly null. Regarding the DRA method, due to the use of a random hash function, some users are not fully satisfied, particularly when the SINR threshold is high.

The same observation can be made in high-density networks. In fact, as shown in Fig. 7, with Q-FCRA and FCRA, more than 80% and 70% of femtocells, respectively, have their TSR above 0.9, whereas C-DFP and DRA fail to provide enough satisfaction, and only 30% of femtocells are able to achieve this rate. This is due to the high number of constraints for C-DFP in high network density and the use of a random hashing function for DRA, which results in performance degradation.

Let us focus on the performance of Q-FCRA in terms of TSR for the 200-node network case and under a high interference level (i.e., $SINR = 25$ dB). Fig. 8(a) and (b) shows the rate of clusters and the average TSR per cluster, respectively. We can observe that most of the clusters contain between two and seven femtocells, and the largest cluster is formed by ten nodes, as $CS_{\text{th}} = 10$ in our simulations [see Fig. 8(a)]. This allows our approach to converge to the optimal solution in a timely manner since, according to Fig. 6, the optimal solution is guaranteed if the number of nodes does not reach the limit of 20. Moreover, notice that generating the optimal solution does not imply as necessary a full satisfaction rate. Indeed, we can see in Fig. 8(b) that the average TSR per cluster decreases with the increase in the cluster size since in this case, the traffic demands per cluster increase and can exceed the 100 available resources. Note that the corresponding values lie between 0.7 and 1 for all clusters, which confirms the results obtained in Fig. 7(b).

Fig. 9(a) further investigates how femtocells' interference degree is taken into account, illustrating the mean TSR as a function of the interference degree for the same scenario (i.e.,

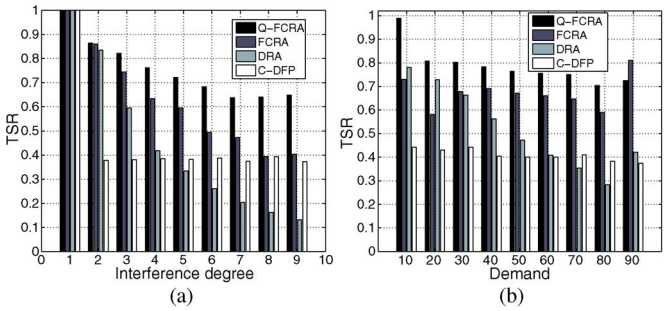


Fig. 9. TSR distribution in high-density networks, i.e., $SINR = 25$ dB. (a) Mean TSR per interference degree. (b) Mean TSR per demand.

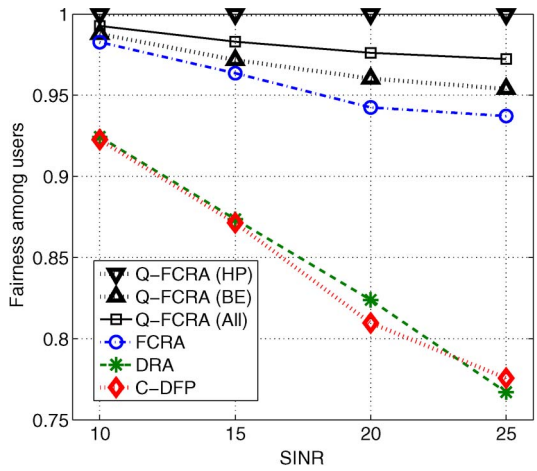


Fig. 10. Fairness comparison.

200-FAPs and 25-dB SINR threshold case). We can see that Q-FCRA is better performing for all interference degrees. In fact, while C-DFP remains around 40%, DRA exponentially decreases, reaching 10% in high-interference degree. In addition, to assess how the allocated resources are affected by the demand volume, Fig. 9(b) plots the mean TSR as a function of the femtocell demand. Globally, C-DFP shows a roughly constant behavior, which implies that its resource allocation is done irrespective of the demands. On the other hand, DRA decreases with growing demands, thus penalizing femtocells with higher demands. However, both FCRA and Q-FCRA approaches try to fairly satisfy all users, even for high demands.

3) Fairness: Regarding the given results, it is important to assess if resources are fairly distributed between users. Fig. 10 shows Jain's fairness index calculated as the average for all the networks. Note that in the best case, it is equal to 1, and it is reached when all users receive the same allocation. We distinguish here between HP and BE users in the Q-FCRA scheme and then show the average for all users. Since for HP users, the distribution highly privileges full satisfaction, unsatisfied users will receive zero resources, thus decreasing the fairness for HP users. However, for BE users, even for the worst case scenario (i.e., high interference level and high-density networks), the total fairness index is better with Q-FCRA reaching approximately 0.95 compared with 0.77 for C-DFP and DRA. Although FCRA uses the same algorithm for BE users allocation, the improvement over FCRA is due to the fact that with Q-FCRA, the reallocation of unused resources by

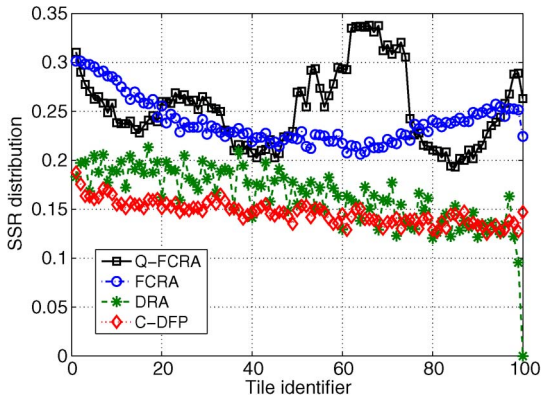


Fig. 11. SSR per tile distribution in high-density networks, i.e., $SINR = 25$ dB.

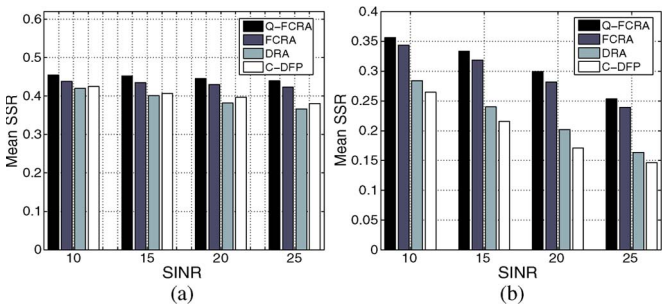


Fig. 12. Mean SSR versus SINR. (a) Low-density networks. (b) High-density networks.

rejected HP users allows a better distribution for the remaining BE users.

4) SSR: Fig. 11 investigates how each tile is reutilized in the network, illustrating the reuse rate of each tile k $(1 \leq k \leq 100)$ for the 200-FAP network case with $SINR = 25$ dB. We can see that tiles reuse cannot exceed 20% for both DRA and C-DFP. However, our approaches perform better particularly for Q-FCRA, where the SSR metric can reach 35% for some segment of tiles. This means that Q-FCRA enhances the SSR up to a factor of 1.75 compared with DRA and C-DFP.

Fig. 12 plots the mean SSR of the underlying schemes as a function of SINR for both low- and high-density networks. Two main observations can be made. First, both Q-FCRA and FCRA offer the highest SSR values, particularly in the case of high-density networks, where the gain of Q-FCRA over FCRA, DRA, and C-DFP can attain, respectively, 2%, 10%, and 13% [see Fig. 12(b)]. This means that Q-FCRA enhances the SSR by a factor of 1.07, 1.5, and 1.76, on average, compared with FCRA, DRA, and C-DFP, respectively. Second, we can notice that the SSR metric decreases with the increase in SINR for all the strategies since the interference degree of each FAP increases. This is clearly shown in Fig. 12(b). Indeed, in this case and according to our approach, fewer but more populated clusters are formed. This results in decreasing the possibility of reutilization of the same tile among the constructed clusters. Recall that our schemes do not allow the utilization of the same tile within the same cluster.

5) Computation and Convergence Time Analysis: Last but not least, it is important to assess if the overall good performances of Q-FCRA come at the expense of higher time complexity compared with other schemes. Table I reports the computation and convergence time needed for Q-FCRA, FCRA, and C-DFP methods to solve the resource-allocation problem. Note that C-DFP does not include a convergence time, since it has a global view of the network.

TABLE I
COMPUTATION AND CONVERGENCE TIME (IN SECONDS) OF Q-FCRA, FCRA, AND C-DFP METHODS, I.E., $SINR = 15$ dB

Network size	50	100	150	200
FCRA	0.013 ± 0.003	0.022 ± 0.005	0.03 ± 0.012	0.07 ± 0.01
Q-FCRA	0.09 ± 0.02	0.12 ± 0.03	0.22 ± 0.03	0.30 ± 0.04
C-DFP	1.59 ± 1.0	5.56 ± 0.58	6.52 ± 0.04	6.80 ± 0.09

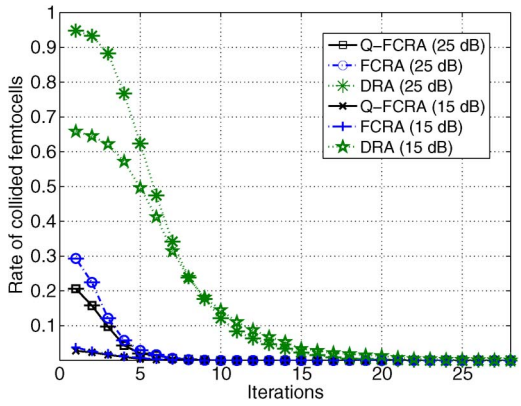


Fig. 13. Convergence time of Q-FCRA, FCRA, and DRA methods in high-density networks.

From this table, we can observe that FCRA, serving only BE users, always converges in few milliseconds. Q-FCRA, on the other hand, needs a little more time to converge, since it executes the algorithm in two steps to deliver the solution for both HP and BE users. However, in both schemes, the computation and convergence time remains very low, i.e., below 0.03 s for the worst case scenario with a high-density network, whereas for C-DFP, the computation time exponentially grows with the network size, which shows the efficiency of our method. Note that for both Q-FCRA and FCRA, the reported time in that table corresponds to the mean value of the time needed for all constructed clusters to resolve the resource-allocation problem.

This is further evidenced in Fig. 13, where we plot the convergence time regarding the high-density network case (i.e., the 200-FAP case) and using different values of SINR. The X-axis represents the number of necessary control frames sent by the end users to their associated FAPs when collisions occur. The Y-axis represents the portion of FAPs that experienced a collision on one of their allocated tiles. Note that in each simulation, we vary the FAP network topology, the number of end users associated with each FAP, and their traffic demands (in terms of requested tiles). In that figure, we can notice that both Q-FCRA and FCRA converge to a stationary allocation within ten frames. On the other hand, DRA needs almost 21 frames to converge. This is related to the totally distributed nature of DRA, which can be interpreted as a negative effect as this can increase the rate of collided FAPs in the whole network due to the absence of coordination between FAPs.

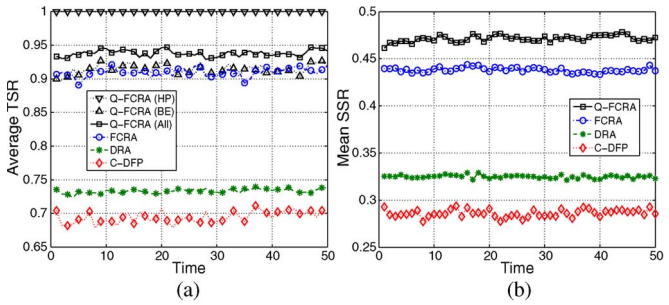


Fig. 14. Impact of users' mobility on TSR and SSR, i.e., $SINR = 25$ dB. (a) TSR. (b) SSR.

B. Mobile Environment Scenario

In this scenario, we study the dynamics of users' connections considering the variation in time of their positions, demands as well as the network load and the SD. Due to space limitations, the results reported here only concern high-density femtocell networks. This section is divided into three parts, where we investigate the impact of each of the aforementioned parameters (i.e., users' mobility rate, network load, and SD) on the relative performance of all studied strategies.

1) Impact of Users' Mobility Rate: The impact of users' mobility on the TSR and the SSR is shown in Fig. 14. In this experiment, the users' mobility is supposed to be high (see Fig. 1), $SINR = 25$ dB, and epoch δ_t is computed according to (2) with $1/\mu = 10$ time unit, $p_3 = p_6 = 0.3$ and $p_7 = 0.6$ (i.e., medium network load and medium SD).

We can observe in Fig. 14(a) that the TSR of accepted HP users is obviously equal to one with Q-FCRA, whereas BE users receive almost the same TSR with either Q-FCRA or FCRA. On average, the overall TSR with Q-FCRA (respectively, FCRA) slightly varies between 92% and 96% (respectively, 88% and 92%), whereas for the other two schemes, it is below 75%. The same reasoning holds when analyzing the impact of users' mobility on the SSR metric [see Fig. 14(b)]. Indeed, we can observe that the SSR metric remains almost constant over time, and the gain of Q-FCRA over FCRA, DRA, and C-DFP can attain in this case 3%, 14%, and 19% on average, respectively. This confirms the robustness of our strategies and can be explained in twofold. First, our mobility-aware resource-allocation process is performed every epoch δ_t, which takes into account the mobility behavior of femto users. Second, the clustering updating process, which is performed only if the interference caused by a user movement increases above a predefined threshold (i.e., 5 dB in our case), allows less frequent adjustment. However, we note that the C-DFP performance is more time sensitive than the other schemes since, in this case, the resource-allocation process does not take into account the interference caused by users' mobility.

2) Impact of Network Load: Now, we study the impact of network load on the system performance. The network load indicates the density of the simultaneously attached users to their FAPs that have active communication sessions. According to our indoor mobility model, a user becomes active when its status changes from state "O" to state "S." Recall that in our simulations, users are uniformly distributed within the residence with a maximum number of ten users per FAP and are initially at state "O." Hence, to control this parameter, we consider three values of probability p_7, as shown in Table II. This experiment is achieved while considering medium users' mobility rate and medium SD.

TABLE II
USERS' NETWORK LOAD PARAMETERS

Network load	Low (30%)	Medium (60%)	High (90%)
p_7	0.3	0.6	0.9

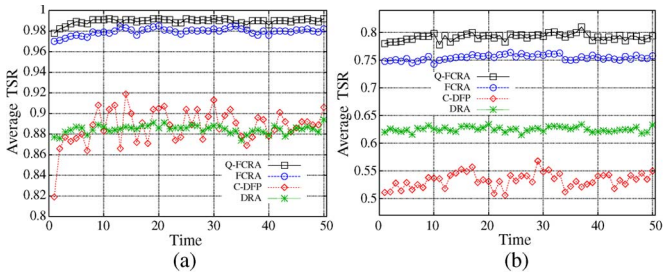


Fig. 15. Impact of network load on TSR. (a) Low load, i.e., $SINR = 10$ dB. (b) High load, i.e., $SINR = 25$ dB.

Fig. 15 shows the impact of network load on TSR for all schemes in the 200-FAP network case and under two different values of SINR, i.e., 10 and 25 dB. We can observe that Q-FCRA outperforms again the other schemes. However, one should note that both FCRA and Q-FCRA almost have the same performance in the low-load and low-interference-level case [see Fig. 15(a)], since there are enough resources per cluster to satisfy the attached users. We also note in Fig. 15(a) that both C-DFP and DRA have the same performance, whereas in a high-load and high-interference scenario, the C-DFP performance degrades below that of DRA, as shown in Fig. 15(b). This can be explained by the high number of constraints that C-DFP needs to take into account in this corresponding scenario. In addition, we can observe that the performance decreases for all strategies when the network load increases, since increasingly more demands need to be satisfied.

3) Impact of SD: Finally, we study the impact of SD on the system performance. However, due to space limitations, we did not include the corresponding curves. The SD determines for each user the average duration before the user ends its session and becomes "passive" (i.e., before a user moves from states "S" or "M" to state "O" of the indoor mobility model in Fig. 1). Hence, this parameter is controlled by varying probabilities p_3 and p_6. Note that this parameter can be expressed as follows:

$$SD = \Pi_S \times 1/\mu + \Pi_M \times 1/\mu = (1 - \Pi_O) \times 1/\mu \quad (10)$$

where Π_i denotes the steady-state probability for state "i" of the Markov chain shown in Fig. 1, and $1/\mu$ is the mean sojourn time of a femto user within its corresponding FAP. Recall that, in our simulations, $1/\mu = 10$ time unit.

Similar behaviors as in Fig. 15 are observed here. Indeed, we notice that Q-FCRA outperforms other strategies. The gain of Q-FCRA over FCRA increases with the increase in the SD. In addition, TSR values for all schemes decrease when the SD increases. This is because the allocated resources are not released until the session expires.

VII. Conclusion

In this paper, we have investigated the joint resource-allocation and admission control problem in OFDMA-based femtocell networks, taking into account users QoS requirements. Two types of users are considered: QoS-constrained HP users and BE users. An HP user can be the femto owner, while the BE users can be the visitors (in open or hybrid access), or HP and BE users are differentiated based on the price they pay for the service. We have proposed a cluster-based hybrid strategy as an alternative to centralized and distributed approaches. Our proposal, which is called Q-FCRA, involves three main stages: 1) cluster formation, 2) intracluster resource allocation and admission control, and 3) intercluster resource contention resolution. In particular, the second stage includes first resource allocation and admission control for HP users by minimizing the sum of corresponding elastic variables. Then, BE users resource allocation is performed by resolving a Min–Max optimization problem. Through extensive simulations, we showed that our approach can achieve significant gains in terms of the rate of rejected users in the network, fairness of the system, users TSR, and SSR, compared with those used as benchmarks (i.e., FCRA, DRA, and C-DFP). Specifically, we showed that our approach can reject 20% of users' demands in high-density networks with a high interference level, whereas both DRA and C-DFP reject more than 40%. In addition, we showed that our approach allows more than 80% of femtocells to have their TSR above 0.9, whereas C-DFP and DRA fail to provide enough satisfaction, and only 30% of femtocells are able to achieve this rate. Moreover, we showed that Q-FCRA enhances the SSR by a factor of 1.5 and 1.75, on average, compared with DRA and C-DFP, respectively. Finally, we demonstrated that Q-FCRA converges more quickly than DRA and has low computation time compared with C-DFP. These results make our approach an efficient solution for resource allocation in femtocell networks.

References

[1] A. Hatoum, R. Langar, N. Aitsaadi, R. Boutaba, and G. Pujolle, "Q-FCRA: Qos-based OFDMA femtocell resource allocation algorithm," in *Proc. IEEE ICC*, 2012, pp. 5151–5156.

[2] I. Guvenc, M.-R. Jeong, F. Watanabe, and H. Inamura, "Hybrid frequency assignment for femtocells and coverage area analysis for cochannel operation," *IEEE Commun. Lett.*, vol. 12, no. 12, pp. 880–882, Dec. 2008.

[3] V. Chandrasekhar and J. Andrews, "Power control in two-tier femtocell networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 8, pp. 4316–4328, Aug. 2009.

[4] Y. Kim, S. Lee, and D. Hong, "Performance analysis of two-tier femtocell networks with outage constraints," *IEEE Trans. Wireless Commun.*, vol. 9, no. 9, pp. 2695–2700, Sep. 2010.

[5] J.-H. Yun and G. Shin, "CTRL: A self-organizing femtocell management architecture for co-channel deployment," in *Proc. Int. Conf. MobiCom*, 2010, pp. 61–72.

[6] S.-Y. Lien, Y.-Y. Lin, and K.-C. Chen, "Cognitive and game-theoretical radio resource management for autonomous femtocells with QoS guarantees," *IEEE Trans. Wireless Commun.*, vol. 10, no. 7, pp. 2196–2206, Jul. 2011.

[7] N. Chakchouk and B. Hamdaoui, "Uplink performance characterization and analysis of two-tier femtocell networks," *IEEE Trans. Veh. Tech.*, vol. 61, no. 9, pp. 4057–4068, Nov. 2012.

[8] D. Lopez-Perez, A. Valcarce, G. de la Roche, and J. Zhang, "OFDMA Femtocells: A roadmap on interference avoidance," *IEEE Commun. Mag.*, vol. 47, no. 9, pp. 41–48, Sep. 2009.

[9] H. Lee and Y. L. D. C. Oh, "Mitigation of inter-femtocell interference with adaptive fractional frequency reuse," in *Proc. IEEE ICC*, Jun. 2010, pp. 1–5.

[10] K. Sundaresan and S. Rangarajan, "Efficient resource management in OFDMA femtocells," in *Proc Int. Symp. Mobihoc*, 2009, pp. 33–42.

[11] V. Chandrasekhar and J. Andrews, "Spectrum allocation in tiered cellular networks," *IEEE Trans. Commun.*, vol. 57, no. 10, pp. 3059–3068, Oct. 2009.

[12] L. G. U. Garcia, K. L. Pedersen, and P. E. Mogensen, "Autonomous component carrier selection: Interference management in local area environments for LTE-advanced," *IEEE Commun. Mag.*, vol. 47, no. 9, pp. 110–116, Oct. 2009.

[13] Y. Arslan, J. Yoon, K. Sundaresan, V. Krishnamurthy, and S. Banerjee, "A femtocell resource management system for interference mitigation in OFDMA networks," in *Proc. Int. Conf. MobiCom*, 2011, pp. 25–36.

[14] M. Rahman and H. Yanikomeroglu, "Enhancing cell-edge performance: A downlink dynamic interference avoidance scheme with inter-cell coordination," *IEEE Trans. Wireless Commun.*, vol. 9, no. 4, pp. 1414–1425, Apr. 2010.

[15] G. de la Roche, A. Valcarce, D. Lopez-Perez, and J. Zhang, "Access control mechanisms for femtocells," *IEEE Commun. Mag.*, vol. 48, no. 1, pp. 33–39, Jan. 2010.

[16] H.-S. Jo, P. Xia, and J. G. Andrews, "A downlink femtocell networks: Open or closed?" in *Proc. IEEE ICC*, Jun. 2011, pp. 1–5.

[17] S. Yun, Y. Yi, D. Cho, and J. Mo, "Open or close: On the sharing of femtocells," in *Proc. IEEE INFOCOM*, 2011, pp. 116–120.

[18] A. Hatoum, N. Aitsaadi, R. Langar, R. Boutaba, and G. Pujolle, "FCRA: femtocell cluster-based resource allocation scheme for OFDMA networks," in *Proc. IEEE ICC*, 2011, pp. 1–6.

[19] G. Fodor and P. Skillermark, "Performance analysis of a reuse partitioning technique for multi-channel cellular systems supporting elastic services," *Int. J Comm. Syst.*, vol. 22, no. 3, pp. 307–342, Mar. 2009.

[20] M. Ergen, S. Coleri, and P. Varaiya, "QoS aware adaptive resource allocation techniques for fair scheduling in OFDMA based broadband wireless access systems," *IEEE Trans. Broadcast.*, vol. 49, no. 4, pp. 362–370, Dec. 2003.

[21] Y.-J. Choi, J. Kim, and S. Bahk, "QoS-aware selective feedback and optimal channel allocation in multiple shared channel environments," *IEEE Trans. Wireless Commun.*, vol. 5, no. 11, pp. 3278–3286, Nov. 2006.

[22] M. Pischella and J.-C. Belfiore, "Resource allocation for QoS-aware OFDMA using distributed network coordination," *IEEE Trans. Veh. Technol.*, vol. 58, no. 4, pp. 1766–1775, May 2009.

[23] Y.-S. Liang, W.-H. Chung, G.-K. Ni, I.-Y. Chen, H. Zhang, and S.-Y. Kuo, "Resource allocation with interference avoidance in ofdma femtocell networks," *IEEE Trans. Veh. Tech.*, vol. 61, no. 5, pp. 2243–2255, Jun. 2012.

[24] "Evolved universal terrestrial radio access (E-UTRA) and evolved universal terrestrial radio access network (E-UTRAN); overall description; stage 2," 3rd Generation Partnership Project, Sophia-Antipolis, France, 3GPP Tech. Spec. TS 36.300 ver. 10.5.0, Oct. 2011.

[25] "Evolved universal terrestrial radio access (E-UTRA); FDD Home eNode B (HeNB) radio frequency (RF) requirements analysis," 3rd Generation Partnership Project, Sophia-Antipolis, France, 3GPP Tech. Rep 36.921 ver. 10.0.0 rel. 10.

[26] K.-H. Chiang and N. Shenoy, "A 2-d random-walk mobility model for location-management studies in wireless networks," *IEEE Trans. Veh. Technol.*, vol. 53, no. 2, pp. 413–424, Apr. 2004.

[27] H. Aissi, C. Bazgan, and D. Vanderpooten, "Complexity of the min–max and min–max regret assignment problems," *Oper. Res. Lett.*, vol. 33, no. 6, pp. 634–640, Nov. 2005.

[28] S. Leei, "An enhanced IEEE 1588 time synchronization algorithm for asymmetric communication link using block burst transmission," *IEEE Commun. Lett.*, vol. 12, no. 9, pp. 687–689, Sep. 2008.

[29] [Online]. Available: http://www-01.ibm.com/software/integration/optimization/cplex-optimizer/

[30] G. Brown and G. Graves, "Elastic programming: A new approach to large-scale mixed integer optimization," presented at the ORSA/TIMS Conf., Las Vegas, NV, USA, 1975.

[31] J. W. Chinneck and E. W. Dravnieks, "Locating minimal infeasible constraint sets in linear programs," *INFORMS J. Comput.*, vol. 3, no. 2, pp. 157–168, 1991.

[32] E. Hahne, "Round-robin scheduling for max–min fairness in data networks," *IEEE J. Sel. Areas Commun.*, vol. 9, no. 7, pp. 1024–1039, Sep. 1991.

[33] D. Winner, II, "Winner II Channel Models Part I—Channel Models," WINNER, Munich, Germany, Tech. Rep. IST-4-027756, Sep. 2007.

[34] A. Ladanyi, D. Lopez-Perez, A. Juttner, X. Chuy, and J. Zhang, "Distributed resource allocation for femtocell interference coordination via power minimisation," in *Proc. IEEE GLOBECOM*, 2011, pp. 744–749.

Abbas Hatoum received the Diploma in electrical, electronics, computer, and telecommunications engineering from Lebanese University, Beirut, Lebanon, in 2008 and the M.Sc. and Ph.D. degrees in network and computer science from the University of Pierre and Marie Curie—Paris 6, Paris, France, in 2009 and 2013, respectively.

He is currently an R&D Engineer with Ucopia Communications, France, working on third-generation/fourth-generation (3G/4G) off-loading. His current research interests include resource and mobility management, small cells, and 3G/4G mobile offloading.

Rami Langar (M'10) received the M.Sc. degree in network and computer science from the University of Pierre and Marie Curie—Paris 6, Paris, France, in 2002 and the Ph.D. degree in network and computer science from Télécom ParisTech, Paris, in 2006.

In 2007 and 2008, he was a Postdoctoral Research Fellow with the School of Computer Science, University of Waterloo, Waterloo, ON, Canada. He is currently an Associate Professor with the Laboratoire d'informatique de Paris 6 (LIP6), University of Pierre and Marie Curie—Paris 6. His research interests include mobility and resource management in wireless mesh, vehicular ad hoc and femtocell networks, green networking, cloud radio access networks, performance evaluation, and quality-of-service support.

Nadjib Aitsaadi (M'09) received the Engineer Diploma in computer science from the Higher National School of Computer Science, Algiers, Algeria, in 2005; the M.Sc. degree in computer science and networking from the University of Pierre and Marie Curie—Paris 6, Paris, France, in 2006; and the Ph.D. degree in computer science (with honors) from the Laboratoire d'informatique de Paris 6 (LIP6), University of Pierre and Marie Curie—Paris 6, in 2010.

Since September 2011, he has been an Associate Professor of computer science with the University of Paris-Est Créteil Val de Marne, Créteil, France. He is a member of the Laboratory of Image, Signal and Intelligent Systems (LISSI). From June 2010 to August 2011, he was a Research Fellow with the INRIA–HIPERCOM team.

Raouf Boutaba (F'12) received the M.Sc. and Ph.D. degrees in computer science from the University of Pierre and Marie Curie—Paris 6, Paris, France, in 1990 and 1994, respectively.

He is currently a Full Professor of computer science with the University of Waterloo, Waterloo, ON, Canada. His research interests include network and resource and service management in wired and wireless networks.

Dr. Boutaba is the founding Editor-in-Chief of the IEEE TRANSACTIONS ON NETWORK and SERVICE MANAGEMENT (2007–2010) and has served on the editorial boards of several other journals. He has received several best paper awards and other recognitions, such as the Premiers Research Excellence Award, the IEEE Hal Sobol Award in 2007, the Fred W. Ellersick Prize in 2008, the Joe LociCero and the Dan Stokesbury awards in 2009, and the Salah Aidarous Award in 2012. He is a Fellow of the IEEE and the Engineering Institute of Canada.

Guy Pujolle (SM'13) received the Ph.D. degree in computer science from Paris Dauphine University—Paris IX, Paris, France, in 1975 and the "These d'Etat" degree in computer science from the University of Paris-Sud—Paris XI, Orsay, France, in 1978.

He is currently a Full Professor with the University of Pierre and Marie Curie—Paris 6, Paris; a member of the Institut Universitaire de France since 2009; and a member of The Royal Physiographical Academy of Lund, Sweden. He is the French representative on the Technical Committee on Networking at the International Federation for Information Processing. He is the Editor for the Association for Computing Machinery *International Journal of Network Management* and *Telecommunication Systems* and the Editor-in-Chief of *Annals of Telecommunications*. He is a pioneer in high-speed networking, having led the development of the first Gbit/s network to be tested in 1980. He participated in several important patents such as Deep Packet Inspection, virtual networks, or virtual access points. He is the Cofounder of QoSMOS (www.qosmos.fr), Ucopia Communications (www.ucopia.com), EtherTrust (www.ethertrust.com), Virtuor (www.VirtuOR.fr), and Green Communications (www.greencommunications.fr).