This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2020.3022681, IEEE Transactions on Mobile Computing

Coordinated Framework for Spectrum Allocation and User Association in 5G HetNets with mmWave

Kinda Khawam*, Samer Lahoud†, Melhem El Helou†, Steven Martin‡, Feng Gang§
*University of Versailles, France
†ESIB, Saint Joseph University of Beirut, Lebanon
‡LRI, Paris-Sud University, France
§University of Electronic Science and Technology of China, China

***Abstract*—Dense deployment of small cells operating on different frequency bands based on multiple technologies provides a fundamental way to face the imminent thousand-fold traffic augmentation. This heterogeneous network (HetNet) architecture enables efficient traffic offloading among different tiers and technologies. However, research on multi-tier HetNets where various tiers share the same microwave spectrum has been well-addressed over the past years. Therefore, our work is targeted towards novel multi-tier HetNets with disparate spectrum (microwave and millimeter wave). In fact, despite the huge capacity brought by millimeter-wave technology, the latter will fail to provide universal coverage, especially indoor, and so mmWave will inevitably co-exist with a traditional sub-6GHz cellular network. In this work, we propose a coordinated user association and spectrum allocation by resorting to non-cooperative game theory. In fact, in such an arduous context, efficient distributed solutions are imperative. Extensive simulation results show the precedence of our coordinated approach in comparison with state-of-the-art heuristics. Moreover, we evaluate the impact of various network parameters, such as mmWave density, cell load, and user distribution and density, offering valuable guidelines into practical 5G HetNet design. Finally, we assess the benefit brought by massive MIMO for mmWave in such a highly heterogeneous setting.**

***Index Terms*—5G HetNet, mmWave, User association, spectrum allocation, non-cooperative game theory, convex optimization.**

I. Introduction

It is undeniable that there would be no single technology that can meet the stringent 5G requirements. These requirements consist in achieving more capacity and better Quality of Experience (QoE) while servicing a very large number of wireless connections for both human and machine applications, with diverse characteristics. To address the capacity and data rate demands, current consensus is to aggregate more bandwidths and infrastructure nodes, especially by resorting to network densification and adopting mmWave spectrum [1].

Hence, future wireless networks will remain highly heterogeneous, from a **dual** perspective to deliver the 5G performance expectations. First, we have the heterogeneity in spectrum, since frequency bandwidth below 6 GHz is very crowded and can no longer meet the aggressive requirements in terms of network capacity. Accordingly, the millimeter wave is considered as a good candidate to attain Gigabit communications. Second, we have the unavoidable heterogeneity in cell size where a diverse set of small-cells will still overlay macro-cells. Macro-cells deliver basic long-range coverage, and small cells provide short-range but high quality communication to users in their vicinity. Particularly, the cell size for mmWave spans only a few hundred meters because of the high attenuation in the corresponding frequency bands [2], [3]. Hence, 5G Heterogeneous Networks (HetNet) are typically composed of multiple tiers: macro Base Stations (BSs), with a double overlay of femto BSs, and mmWave BSs. If mmWave BSs are noise rather than interference-limited, the increased density of femto BSs renders co-tier and cross-tier interference prohibitive for traditional sub-6GHz networks. In this strenuous context, novel Radio Resource Management (RRM) should be conceived to mitigate such interference while efficiently associating users with the various technologies and tiers.

This **double-facet** heterogeneity complicates the problem of RRM in 5G HetNets. To tackle both aspects, this paper proposes a coordinated spectrum allocation (SA) and user association (UA). The main issue is the interdependence of SA and UA: on the one hand, to devise adequate spectrum allocation algorithms, we need to know the cell load which is reliant on the user association algorithm. On the other hand, the attractiveness of a given cell for the user association relies on its capacity and hence on the number of subchannels assigned by the spectrum allocation scheme. We implement in this framework algorithms for solving the compound problem in a centralized and distributed fashion. We also compare the devised algorithms with state-of-the-art approaches, and assess the benefit brought by massive MIMO for mmWave. In what follows, we examine related work and highlight our contributions.

A. Existing Work

The user association that has recourse to the effect of range expansion techniques in conjunction with spectrum allocation has been formulated in ([4]–[9]). In particular, an analytical approach for biasing and interference coordination was thoroughly studied in ([7]–[9]). Classical user association compounded with spectrum allocation was studied in ([10]–[13]). In [10], the adopted spectrum allocation was restricted to three pre-defined resource allocation strategies, namely, orthogonal deployment, co-channel deployment, and partially shared deployment. The work in [11] extends the work in

[10] by considering pre-defined strategies (*i.e.*, reuse patterns) when the user association is also optimized. Although the problem is non-convex combinatorial, the authors developed efficient algorithms to compute tight upper bounds using convex relaxation. In [12], authors resort to stochastic geometry to obtain a continuum of users in their objective function; the latter property enables them to jointly optimize spectrum allocation and user association. The optimal solution is put forward when the density of users is low, and near-optimal solution is provided with high users' density. In [13], authors further consider power control and BS operation mode, albeit iteratively. Likewise, in [11] and [14], power control and user association is handled iteratively. The work in [15] investigates also spectrum allocation and user association employing frequency reuse patterns. Yet, spectrum bandwidth of each cell is adapted to its load similarly to [16] in order to cope with spatially inhomogeneous traffic distribution. In fact, the majority of literature does not directly consider the imbalance of cell load in partitioning the frequency bands, and thus the devised algorithms perform poorly unless traffic is uniformly distributed. Hence, three algorithms that alternately evaluate spectrum partition, cell load and user association are applied in [15] until convergence to a fixed point.

The present work will evaluate the impact of integrating mmWave spectrum as a key enabler to achieving gigabyte-level data traffic in future networks. Lately, mmWave frequencies have attracted a lot of attention. Activities for 5G have been launched in 3GPP [17] and completed the practical stage of defining use cases and objectives. In [18], it has been proven that a heterogeneous architecture with mmWave small-cell base stations overlaid on macro cellular network can tremendously increase the system capacity. In [19], a proof-of-concept of such a mmWave overlay HetNet is provided. In particular, a HetNet with commercial LTE and mmWave access is studied. Another important work handling field experimentation is found in [20] where integrated proof-of-concept has been developed for a mmWave-integrated cellular network. This growing interest establishes the key role of mmWave technology in enhancing the capacity of 5G HetNets.

The work in ([21]–[24]) are the closest to ours in evaluating a multi-tier network with multiple technologies using bandwidth below and beyond 6 GHz. However, they only consider user association and overlook spectrum allocation. In particular, the work in [22] considers only mmWave small cells while the work in [24] considers a particular scenario where small cells are assumed to be deployed linearly along roads. Besides, the work in [25] and [26] have similarity with our UA scheme where a weighted proportional fair allocation is adopted. However, the distributed UA in ([25], [26]) stems from an optimal centralized problem by resorting to dual decomposition; whereas, our approach is inherently distributed, built on a game theoretic model. Furthermore, we consider a HetNet that includes mmWave to be in phase with the technological innovation of 5G, contrarily to the mentioned references. More importantly, spectrum allocation is not considered in both [25] and [26].

B. Motivation and contributions

This work presents a coordinated approach for spectrum allocation SA and user association UA in future wireless HetNets. We summarize in this section the major contributions and the demarcation from existing work.

- Contrary to existing work where SA and UA problems are addressed independently, the coordinated framework enables to tailor the spectrum allocation to the user distribution and perform user association accordingly. As a matter of fact, such framework yields load-aware RRM that prevents from over-dimensioning radio resources, or defectively associating users to crowded cells.
- In our work, we devise an original method for estimating the cell load. This method enables to explicit the interdependence between spectrum allocation and user association. In fact, the intricacy between the two problems is typically tackled by iterative procedures ([10]–[13]). These procedures necessitate cumbersome computations and jam the two RRM problems at the same timescale. However, spectrum allocation and user association usually take place on different timescales: SA is a dimensioning task that takes place prior to UA that is more dynamic and frequent. Henceforth, our proposed method enables to obtain a pertinent estimation of any cell load that will serve as a guideline to allocate the spectrum adequately. This estimation takes into account the user distribution and balances between spectral efficiency and spectrum reuse. Afterwards, UA is applied based on the operated spectrum allocation. Finally, we provide insights on how to fine tune the balance between spectral efficiency and spectrum reuse, and hence improve user rates.
- While the spectral resources are disjoint between both types of cells, the user association is still very challenging in this context. The original contribution of our work resides in evaluating the impact of mmWave cells and their density on the coordinated problem of SA and UA in 5G HetNets. Moreover, as already highlighted in the literature [3], we further assess the shortcomings of traditional power-based UA in a heterogeneous setting with sub-6GHz and mmWave technologies. In particular, the work in [23] proposes a different bias per tier and per technology in a sub-optimal two-step user association procedure to address that issue. We validate the difficulty to circumvent the misleading impact of received power in a highly heterogeneous network. Finally, we assess the advantages of massive MIMO in mmWave.
- We introduce in this work original mathematical formulations for the spectrum allocation and the user association problems. First, we formulate SA for LTE femto cells as a non-cooperative game. Femto BSs allocate RBs (Resource Block) in a way to meet the cell load estimation and selfishly strive to minimize perceived interference. We provide a formal mathematical proof that the portrayed game has an exact potential function. Such property guarantees the convergence to pure Nash equilibrium by applying simple Best Response dynamics. Second, we

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2020.3022681, IEEE Transactions on Mobile Computing

tackle user association in 5G networks following two approaches. We formulate a centralized approach and solve it using convex optimization tools. Then, we formulate a distributed approach as a non-cooperative game, where users are players that independently maximize their rate. This game is proven to converge to a unique Nash equilibrium. Owing to the exact potential property, an iterative Best Response algorithm permits attaining such equilibrium.

- We implement in our coordinated framework various spectrum allocation and user association algorithms following centralized, distributed, or basic state-of-the art approaches. Our contribution consists in providing a benchmark where different combinations of algorithms are examined. We provide a thorough analysis for the impact of various network parameters, such as mmWave density, cell load, and user distribution and density. Furthermore, we do not restrain our evaluation to a uniform user distribution and assess the impact of so-called crowd scenario in a given network region.

The rest of the paper is organized as follows. In section II, the network model is thoroughly explained. In section III, our devised coordinated framework is detailed for spectrum allocation and user association. In section IV, we assess the performances of our coordinated approach through various combinations of SA and UA algorithms. By evaluating the impact of different important network metrics, we succeed in offering valuable guidelines into practical 5G HetNet design. In section V, we adapt the network model to include massive MIMO for mmWave and evaluate the corresponding impact on network performances. We finally conclude the paper in section VI.

II. Network Model

In this section, we describe the network topology and radio model used in this work. Although our framework adapts to different deployment scenarios, we select a realistic context for performance evaluation consisting of a multi-tier heterogeneous 5G network. We consider a three-tier wireless network comprising macro LTE cells, femto LTE cells, and mmWave cells.

Tri-sectorized macro-BSs are distributed according to a hexagonal structure with a double overlay of femto BSs and mmWave BSs randomly positioned. We denote by $\mathcal{J}$ the total set of BSs in the network, including $\mathcal{J}^{LTE}$ which is the set of macro BSs and femto BSs, and $\mathcal{J}^{mmW}$ the set of mmWave BSs. We focus on the network downlink where users can concurrently connect to multiple BSs.

As we are dealing with technologies working in disjoint frequency bands, inter-band interference is negligible and SINR expressions can be derived independently as follows.

A. SINR for LTE

We consider that OFDMA (Orthogonal Frequency Division Multiple Access) is used as the multiple access scheme in macro BSs and femto BSs. The time and frequency radio resources are grouped into time-frequency RBs. An RB is the smallest radio resource unit that can be scheduled to a mobile user. Each RB consists of N_s OFDMA symbols in the time dimension and N_f sub-carriers in the frequency dimension (N_s = 7 as in the most used formats and N_f = 12). The set of RBs is denoted by $\mathcal{K}$, and the set of users is denoted by $\mathcal{I}$.

As aforementioned, spectrum allocation consists of computing the allocation of RBs to LTE BSs. The output of such allocation is a set of variables x_{jk} identifying if RB $k \in \mathcal{K}$ is allocated to BS $j \in \mathcal{J}^{LTE}$ as in:

$$x_{jk} = \begin{cases} 1 & \text{if RB } k \text{ is used by BS } j, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

On each allocated RB k, the SINR observed by user $i \in \mathcal{I}$ when connected to LTE BS j, is denoted by $SINR_{ijk}^{LTE}$ and given as:

$$SINR_{ijk}^{LTE} = \frac{p_{jk} g_j g_i \gamma_{ij}}{\sum_{\substack{j' \in \mathcal{J}^{LTE} \\ j' \neq j}} p_{j'k} x_{j'k} g_{j'} g_i \gamma_{ij'} + p_N}. \quad (2)$$

The numerator of (2) represents the received signal power, and the denominator represents the sum of the interfering signals and the thermal noise power per RB denoted by p_N. Interfering signals solely result from the transmission of LTE BSs using the same RBs (for which $x_{j'k} = 1$) since mmWave BSs use disjoint frequency bands.

In the SINR expression, we denote by p_{jk} the transmit power of LTE BS j on RB k, g_j and g_i the BS antenna gain and the user antenna gain respectively, γ_{ij} the pathloss between BS j and user i. Note that we assume a constant power level per RB in the SINR expression. This power level per RB is independent of the number of allocated RBs and is given by $p_{jk} = P_j / |\mathcal{K}|$, where P_j is the total power of BS j and $|\mathcal{K}|$ the total number of RBs. As shown in [27], using a constant power allocation averages the impact of interference and does not hinder the system performance in a multi-cell network.

B. SINR for mmWave

For mmWave, we assume that the totality of the spectrum is used in each cell. Particularly, the coverage of mmWave cells is very limited because of the high attenuation in the corresponding frequency bands. Thus, a full reuse of the spectrum does not generate harmful interferences. As a result, we assume that the totality of spectrum is used in each mmWave BS. It is then unavailing to consider the RB granularity, especially if it leads to loss in generality. In fact, our model goal is to encompass all possibilities open to 5G/mmWave specifications. Hence, our model remains valid and pertinent:

- For different 5G NR (new radio) numerologies (i.e., different RB definitions), and,
- For different mmWave technologies, even if they do not use OFDM (e.g., IEEE 802.11ad uses single-carrier modulation scheme).

The SINR observed by user $i \in \mathcal{I}$, when connected to mmWave BS $j \in \mathcal{J}^{mmW}$, can then be expressed as:

$$SINR_{ij}^{mmW} = \frac{p_j g_j g_i \gamma_{ij}}{\sum_{\substack{j' \in \mathcal{J}^{mmW} \\ j' \neq j}} p_{j'} g_{j'} g_i \gamma_{ij'} + \tilde{p}_N} \quad (3)$$

The numerator of (3) represents the received signal power, and the denominator the sum of the interfering signals and the thermal noise power in the totality of the spectrum denoted by $\tilde{p}_N$. Interfering signals result from the transmission of all mmWave BSs, as in a frequency reuse-1.

In the SINR expression, we denote by p_j the transmit power of mmWave BS j, γ_{ij} the pathloss between BS j and user i. We note that the pathloss model is different from that of LTE BSs: mmWave BSs and LTE BSs operate on different carrier frequencies, have different antenna heights, cover and serve different environment types, and thus undergo different attenuations. In our work, we implement the mmWave pathloss model as given in the seminal work in [28].

C. Peak Rate

The radio conditions of a mobile user yield the instantaneous peak rate it can obtain when connected alone to a given BS. Those conditions are assumed to be invariant as neither mobility nor fading are taken into account. Thus, the peak rate ρ_{ij} of user $i \in \mathcal{I}$ when associated with BS $j \in \mathcal{J}$ is a function of the perceived SINR as in:

$$\rho_{ij} = \begin{cases} W_j f_j(SINR_{ij}^{mmW}) & \text{if } j \in \mathcal{J}^{mmW}, \\ \sum_k W_k f_j(SINR_{ijk}^{LTE}) x_{jk} & \text{if } j \in \mathcal{J}^{LTE}, \end{cases} \quad (4)$$

where W_j and W_k are respectively the cell and RB bandwidth, and f_j is a function that computes the spectral efficiency for a given SINR. This function takes into account the adaptive modulation and coding in LTE and is implemented according to the technical specification in [29]. For mmWave, we make use of the Shannon formula as an upper bound of the spectral efficiency[1].

III. Framework of the Coordinated Spectrum Allocation and User Association Problems

A. Coordinated Problem Formulation

We introduce a mathematical formulation of the coordinated spectrum allocation SA and user association UA problem. We consider a network utility for the downlink of a mutlitier heterogeneous network. While conventional UA basically uses the max-SINR rule, it is evident that max-SINR is inappropriate as it may deprive bad channel quality users from accessing radio resources. Hence, in this work, we strike a good compromise between efficiency and load balancing.

We denote by θ_{ij} the user association variable indicating the percentage of time user i is associated with BS j. The network utility is given by:

$$\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} \theta_{ij}\log(\rho_{ij}) - \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} \theta_{ij}\log(\sum_{i\in\mathcal{I}}\theta_{ij}) \quad (5)$$

When maximizing the network utility, the first term $\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}\theta_{ij}\log(\rho_{ij})$ enables to maximize the user traffic allocations on BSs offering the highest peak rates. Note that taking the logarithm of the peak rate ρ_{ij} prevents allocations on BSs that do not ensure coverage ($\rho_{ij} = 0$). The second term corresponds to the load entropy of the network. In fact, $\Theta_j = \sum_{i\in\mathcal{I}}\theta_{ij}$ corresponds to the load of BS j. Thus, the second term can be rewritten as $-\sum_{j\in\mathcal{J}}\Theta_j\log(\Theta_j)$ and maximizing the entropy leads to balancing the load on the different BSs in the network. Therefore, load balancing avoids congestion on BSs offering the highest peak rates as implied by the first term.

Given (4), the peak rate for LTE cells depends on the spectrum allocation variables x_{jk}, conversely to mmWave cells where the totality of spectrum is allocated in each cell. In the following, the utility function is rewritten as:

$$\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} \theta_{ij}\log(\frac{\rho_{ij}}{\sum_{i\in\mathcal{I}}\theta_{ij}}) \quad (6)$$

The coordinated spectrum allocation and user association is formulated as a mathematical optimization:

$$\begin{aligned}
\underset{\boldsymbol{\theta},\boldsymbol{x}}{\text{maximize}} \quad & \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}^{LTE}} \theta_{ij}\log(\frac{\rho_{ij}(x_{jk})}{\sum_{i\in\mathcal{I}}\theta_{ij}}) && \text{(7a)} \\
& + \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}^{mmW}} \theta_{ij}\log(\frac{\rho_{ij}}{\sum_{i\in\mathcal{I}}\theta_{ij}}) \\
\text{subject to} \quad & \sum_{j\in\mathcal{J}} \theta_{ij} \le 1,\ \forall i \in \mathcal{I}, && \text{(7b)} \\
& \theta_{ij} \ge 0,\ \forall i \in \mathcal{I}, \forall j \in \mathcal{J}, && \text{(7c)} \\
& x_{jk} \in \{0,1\},\ \forall j \in \mathcal{J}^{LTE}, \forall k \in \mathcal{K}. && \text{(7d)}
\end{aligned}$$

The objective function (7a) maximizes the total network utility and ensures proportional fairness between users. Constraints (7b) ensure that each user shares its time on the available BSs.

For clarity, all variables and parameters are summarized in Table I.

$\mathcal{J} = \mathcal{J}^{LTE} \cup \mathcal{J}^{mmW}$	Set of BSes, where $\mathcal{J}^{LTE}$ is the set of LTE BSs and $\mathcal{J}^{mmW}$ is the set mmWave BSs
$\mathcal{I}$	Set of users
$\mathcal{K}$	Set of RBs
θ_{ij}	percentage of time user i is associated with BS j
x_{jk}	RB allocation variable indicating if RB k is alloted to LTE BS j

TABLE I: Sets, parameters and variables in the document

The optimization problem formulated in (7) is a Mixed Integer Non-Linear Program (MINLP). Particularly, the objective function (7a) is non-convex and the problem is intractable for a large number of variables in realistic scenarios.

Recently, in the literature, iterative approaches are still used to overcome the complexity of the original problem in [11], [14]. These approaches consist in iteratively solving the SA and UA problems. These procedures necessitate cumbersome computations and wrongly align the two problems at the same timescale. In fact, SA is a dimensioning task that takes precedence in time over UA that is more frequent.

In this section, we introduce a framework that operates in two levels in order to solve the spectrum allocation and user association problem formulated in (7). These two levels are astutely coordinated in order to provide an efficient solution to the original problem. Precisely, we devise a pertinent estimation of any cell load that will serve as a guideline to

[1]This is an upper bound of upcoming technical specification

allocate the spectrum adequately. This estimation takes into account the user distribution and balances between spectral efficiency and spectrum reuse. Afterwards, UA is applied based on the operated spectrum allocation. Contrarily to existing approaches, our framework is suited to the multiplicity of timescales and takes into account the interdependence of the two problems without having recourse to iterative procedures.

Our framework is represented in two building blocks, as shown in Figure 1. The first building block aims at computing the spectrum allocation for the network cells. We consider separately LTE cells and mmWave cells, as they use disjoint frequency bands. For LTE cells, we start by deciding of the macro-femto spectrum sharing. Then, we allocate the appropriate RBs per cell for each tier, based on a devised cell load estimation introduced in the following section.

The second building block shown in Figure 1 includes peak rate, SINR computation and user association. Considering the spectrum allocation, we compute the user peak rate and SINR per BS for each user in the network. Such computation takes into account the user radio conditions in terms of pathloss and interference. Then, UA is performed following centralized, distributed, and basic state-of-the art approaches.

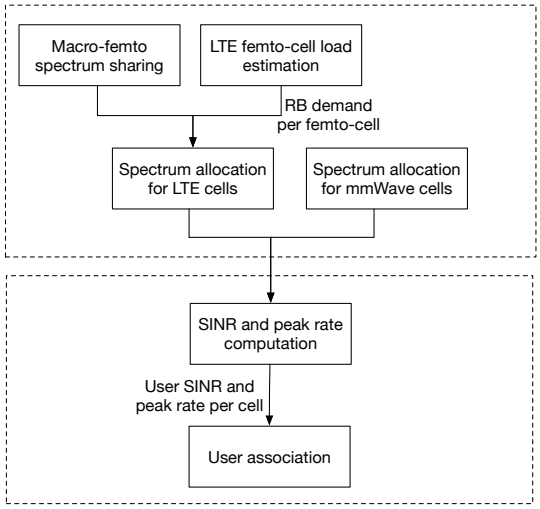

Fig. 1: Building blocks of the SA and UA coordinated framework

We implement in our coordinated framework various SA and UA algorithms. The common framework introduced in this work enables to provide deep understanding of the different building blocks and their mutual dependence. Additionally, it constitutes a scientific benchmark where different combinations of the devised algorithms are examined.

B. Spectrum Allocation

In the downlink, OFDM allows assigning frequency sub-carriers to users within each cell in an orthogonal manner, each sub-carrier having a much lower bandwidth than the coherence bandwidth of the channel. With the use of cyclic prefix insertion, intra-cell interference can be eradicated. However, inter-cell interference remains problematic and must be tackled.

For mmWave, we assume that the totality of spectrum is used in each cell. Particularly, the coverage of mmWave cells is very limited because of the high attenuation in the corresponding frequency bands. Thus, a full reuse of spectrum does not generate harmful interferences.

For LTE cells, OFDMA is very attractive as it enjoys high spectral efficiency and immunity to both frequency selective fading and inter-symbol interference (ISI). The latter encourages the use of frequency reuse-1. However, when the same RB is used in neighboring cells, high interference may occur. The latter can degrade the SINR perceived by serviced users, especially in femto cells that can be randomly close to each other. Consequently, SA starts by allocating the adequate bandwidth to LTE cells as follows:

- We start by estimating the load of each LTE tier based on the maximum received power. A user $i \in \mathcal{I}$ is estimated to be associated with BS j if $P_{ij} \geq P_{ij'}, \forall j' \in \mathcal{J}^{LTE}$, where $P_{ij} = p_j g_j g_i \gamma_{ij}$ is the power received by user i from BS j. Accordingly, we denote by n_j the number of users associated with BS j. Finally, we denote by $n_{femto} = \sum_j n_j, j \in \mathcal{J}^{LTE}_{femto}$ and by $n_{macro} = \sum_j n_j, j \in \mathcal{J}^{LTE}_{macro}$ the load of femto and macro tiers respectively.
- Each LTE tier is allocated a number of RBs proportional to its load: the number of RBs allocated to the macro tier is consequently equal to $\frac{n_{macro}}{n_{macro}+n_{femto}} \times |\mathcal{K}|$. Similarly, the number of RBs allocated to the femto tier is $\frac{n_{femto}}{n_{macro}+n_{femto}} \times |\mathcal{K}|$, where $\mathcal{K}$ is the set of RBs.

Then, in each tier, the RB allocation is done as follows:

- The totality of RBs available for macro cells is deployed with frequency reuse-1, as advocated for increased spectral efficiency.
- The spectrum share available for femto cells is allocated in two steps: first, each femto cell is allocated a number of RBs proportional to its load, according to the load estimation method presented in Section III-B1. Second, the exact RBs are selected in a way to mitigate the co-tier interference resulting from a random BS deployment. The latter RB allocation method is portrayed as a non-cooperative game and presented in Section III-B2.

1) Femto Cell Load Estimation: In our cell load estimation, we allocate a number of RBs to each femto cell proportionally to its load. Computing an exact value of this load necessitates knowing how many users are serviced by each femto. However, such information is not available at the spectrum allocation level as we are in the initial phase of a two-level approach. For this reason, we introduce a pertinent load estimation that captures the user distribution and radio coverage.

Let us consider a femto BS $j \in \mathcal{J}^{LTE}_{femto}$, where $\mathcal{J}^{LTE}_{femto}$ is the set of femto LTE BSs. Given the pathloss model, a user $i \in \mathcal{I}$ is estimated to be associated with BS j if $\gamma_{ij} \geq \gamma_{ij'}, \forall j' \in \mathcal{J}^{LTE}_{femto}$. We denote by L_j the absolute load of BS j or equivalently the number of users associated with this BS. We define a per-cell cluster C_j of BS j as the set of BSs $j' \in C_j$ where $\gamma_{jj'} \geq \gamma_{threshold}$. We consider that the totality of spectrum is used in each per-cell cluster $C_j, \forall j \in \mathcal{J}^{LTE}_{femto}$. Then, the threshold is defined in a way to balance between interference and spectrum reuse. A large threshold reduces the

size of each per-cell cluster enabling higher spectrum reuse at the cost of higher co-tier interference and lower spectral efficiency, and vice versa. Given the notion of per-cell cluster, the relative load of BS j is computed by $l_j = \frac{L_j}{\sum_{j' \in C_j} L_{j'}}$. Finally, the number of RBs allocated to BS j is proportional to its relative load $N_j = l_j \times |\mathcal{K}|$, where $\mathcal{K}$ is the set of RBs.

2) Femto Cell RB Allocation: Inter-cell interference remains a central issue for the design of 5G networks with dense small cells that are particularly vulnerable to interference [30]. In this work, we rely on non-cooperative game theory to devise a distributed RB allocation algorithm. In fact, non-cooperative game theory models the interactions between players competing for a common resource. Hence, it is well adapted to model the spectrum allocation between selfish competing femto cells. Our algorithm builds on the number of RBs computed at the cluster level in the previous section III-B1. Accordingly, each femto cell identifies the pool of the least interfered RBs. Our distributed algorithm minimizes the interference level by only making use of local information available at the femto cells through Channel Quality Indication (CQI) feedbacks [29].

a) RB Allocation Game: We define a multi-player game $\mathcal{G}^{SA}$ between all femto cells. Femto cells are assumed to make their decisions without knowing the decisions of each other. The formulation of the game $\mathcal{G}^{SA} = \left\langle \mathcal{J}^{LTE}_{femto}, S^{SA}, I \right\rangle$ can be described as follows:

- The finite set of femto cells $\mathcal{J}^{LTE}_{femto}$;
- Each femto cell j has to pick N_j RBs among the $|\mathcal{K}|$ available RBs. The action of femto cell j is $\boldsymbol{x}_j = (x_{j1}, ..., x_{j|\mathcal{K}|})$. A *strategy profile* $\boldsymbol{x} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_{|\mathcal{J}^{LTE}_{femto}|})$ specifies the strategies of all players;
- For each femto cell j, the space of pure strategies is S^{SA}_j given by what follows:

$$S^{SA}_j = \{\boldsymbol{x}_j \in \{0,1\}^{|\mathcal{K}|}, \text{such as } \sum_{k \in \mathcal{K}} x_{jk} = N_j,\ \forall k \in \mathcal{K}\}$$

and $S^{SA} = S^{SA}_1 \times ... \times S^{SA}_{|\mathcal{J}^{LTE}_{femto}|}$ is the set of all strategies;
- A set of cost functions $I = (I_1(\boldsymbol{x}), I_2(\boldsymbol{x}), ..., I_{|\mathcal{J}|^{LTE}_{femto}}(\boldsymbol{x}))$ that quantify players' objective for a given strategy profile $\boldsymbol{x}$, where $I_j(\boldsymbol{x})$ is the cost function of femto cell j.

Our RB allocation game enables each femto cell to select a pool of least interfered RBs. Therefore, the cost function must capture the co-tier interference endured by users of a given femto cell. The latter is the harmful signal received from neighboring cells using the same RBs. Computing co-tier interference necessitates knowing the users that are serviced by each femto. However, user association information is not available at the spectrum allocation level. For this reason, we devise a cost function I_j that considers the impact of neighboring femto cells on cell j itself instead of the impact on users associated to femto cell j. This approximation is pertinent as the size of femto cells is small. Accordingly, the co-tier interference endured by femto cell j on RB k can be written as $\sum_{\substack{j' \in \mathcal{J}^{LTE}_{femto},\\ j' \neq j}} x_{j'k} p_{j'k} g_{j'} g_j \gamma_{jj'} + p_N$, where $\gamma_{jj'}$ denotes the pathloss between BS j' and BS j. Then, the cost function of femto j that enables to minimize the total co-tier interference is casted as:

$$I_j = \sum_{k \in \mathcal{K}} x_{jk} \Big(\sum_{\substack{j' \in \mathcal{J}^{LTE}_{femto},\\ j' \neq j}} G_{jj'} \cdot x_{j'k} + p_N \Big), \tag{8}$$

where $G_{jj'} = p_{j'k} g_{j'} g_j \gamma_{jj'}$. The latter is independent of k as the power per RB p_{jk} is constant.

b) Reaching Pure Nash Equilibriums: In a non-cooperative game, an efficient solution is obtained when all players adhere to a Nash Equilibrium (NE). A NE is a profile of strategies in which no player will profit from deviating its strategy unilaterally. Hence, it is a strategy profile where each player's strategy is an optimal response to the other players' strategies:

$$I_j(\boldsymbol{x}_j, \boldsymbol{x}_{-j}) \leq I_j(\boldsymbol{x}'_j, \boldsymbol{x}_{-j}), \forall j \in \mathcal{J}^{LTE}_{femto}, \forall \boldsymbol{x}'_j \in S^{SA}_j, \tag{9}$$

where $\boldsymbol{x}_{-j}$ denotes the vector of strategies played by all femto cells except femto cell j. We turn to potential games to show the existence of NE.

Exact Potential Games: Potential games form a special class of normal form games where the unilateral change of one player strategy $\boldsymbol{x}_j$ to $\boldsymbol{x}'_j$ results in a change of its cost function that is equal to the change of a so-called potential function $\phi : S^{SA} \rightarrow \mathbb{R}$. A potential game [31] admits at least one pure NE which is a desired property.

Proposition 3.1: The game $\mathcal{G}^{SA}$ is an exact potential game. A candidate for the potential function which maps a profile $\boldsymbol{x} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_{|\mathcal{J}^{LTE}_{femto}|})$ to a real is the following:

$$\phi(\boldsymbol{x}) = 1/2 \sum_{j \in \mathcal{J}^{LTE}_{femto}} \sum_{k \in \mathcal{K}} x_{jk} \Big(\sum_{\substack{j' \in \mathcal{J}^{LTE}_{femto},\\ j' \neq j}} G_{jj'} \cdot x_{j'k} + p_N \Big)$$

Proof: We prove that if $\boldsymbol{x}$ and $\boldsymbol{x}'$ are two pure profiles which only differ on the strategy of one BS ℓ, then $I_\ell(\boldsymbol{x}_\ell, \boldsymbol{x}_{-\ell}) - I_\ell(\boldsymbol{x}'_\ell, \boldsymbol{x}_{-\ell}) = \phi(\boldsymbol{x}_\ell, \boldsymbol{x}_{-\ell}) - \phi(\boldsymbol{x}'_\ell, \boldsymbol{x}_{-\ell})$ as $G_{jj'} = G_{j'j}$:

$$\begin{aligned}
&2\phi(\boldsymbol{x}) - 2\phi(\boldsymbol{x}') = \\
&\sum_{\substack{j \in \mathcal{J}^{LTE}_{femto},\\ j \neq \ell}} \sum_{k \in \mathcal{K}} x_{jk} \Big(\sum_{\substack{j' \in \mathcal{J}^{LTE}_{femto},\\ j' \neq j,\\ j' \neq \ell}} x_{j'k} G_{jj'} + x_{\ell k} G_{j\ell} + p_N \Big) \\
&- \sum_{\substack{j \in \mathcal{J}^{LTE}_{femto},\\ j \neq \ell}} \sum_{k \in \mathcal{K}} x_{jk} \Big(\sum_{\substack{j' \in \mathcal{J}^{LTE}_{femto},\\ j' \neq j,\\ j' \neq \ell}} x_{j'k} G_{jj'} + x'_{\ell k} G_{j\ell} + p_N \Big) \\
&+ \sum_{k \in \mathcal{K}} x_{\ell k} \Big(\sum_{\substack{j' \in \mathcal{J}^{LTE}_{femto},\\ j' \neq \ell}} x_{j'k} G_{\ell j'} + p_N \Big) \\
&- x'_{\ell k} \Big(\sum_{\substack{j' \in \mathcal{J}^{LTE}_{femto},\\ j' \neq \ell}} x_{j'k} G_{\ell j'} + p_N \Big) \\
&= \sum_{k \in \mathcal{K}} \sum_{\substack{j \in \mathcal{J}^{LTE}_{femto},\\ j \neq \ell}} (x_{jk} G_{j\ell} + p_N) \cdot (x_{\ell k} - x'_{\ell k}) \\
&+ I_\ell(\boldsymbol{x}_\ell, \boldsymbol{x}_{-\ell}) - I_\ell(\boldsymbol{x}'_\ell, \boldsymbol{x}_{-\ell}) \\
&= 2 \cdot (I_\ell(\boldsymbol{x}_\ell, \boldsymbol{x}_{-\ell}) - I_\ell(\boldsymbol{x}'_\ell, \boldsymbol{x}_{-\ell}))
\end{aligned}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2020.3022681, IEEE Transactions on Mobile Computing

□

Best Response Dynamics for RB allocation: A NE is a static concept that often abstracts away the question of how it is reached. Thus, the main challenge in non-cooperative game theory is to devise practical algorithms to reach those equilibriums. The simplest example of such algorithms are repeated Best Response dynamics: each player selects the best (locally optimal) response to other players' strategies, until convergence. However, convergence of Best Response dynamics is not guaranteed in general. In this work, we are in presence of an exact potential game where a greedy Best Response algorithm permits attaining Pure NEs (PNE) according to [32].

The best response strategy of a player is the one that minimizes its cost given other players' strategies. The pseudo-code of our distributed RB allocation problem is presented in Algorithm 1. Our algorithm iterates until convergence to a PNE (Line 8). Note that in our case, convergence is guaranteed regardless of the players initial strategies. Hence, in the first iteration (Line 1), we set the allocation vectors to zero. At each iteration, each femto cell $j \in \mathcal{J}_{femto}^{LTE}$ minimizes its cost function I_j given by (8) in response to the strategies of other femto cells in the previous iteration (Line 5). This consists in solving the following optimization problem:

$$(\mathcal{P}_j(\boldsymbol{x}_{-j})): \quad \min_{\boldsymbol{x}_j}\{\sum_{k\in\mathcal{K}} x_{jk}(\sum_{\substack{j'\in\mathcal{J}_{femto}^{LTE},\\ j'\neq j}} G_{jj'}\cdot x_{j'k}+p_N)\} \tag{10a}$$

$$\text{subject to} \quad \sum_{k\in\mathcal{K}} x_{jk}=N_j, \tag{10b}$$

$$x_{jk}\in\{0,1\},\ \forall k\in\mathcal{K}. \tag{10c}$$

Problem $(\mathcal{P}_j(\boldsymbol{x}_{-j}))$ solved by each femto cell $j \in \mathcal{J}_{femto}^{LTE}$ at each iteration is an Integer Linear Program (ILP). An ILP is typically solved using a branch-and-bound approach based on linear programming. The idea of this approach is to solve Linear Program (LP) relaxations of the ILP and to look for an integer solution by branching and bounding on the decision variables provided by the LP relaxations. Thus, in a branch-and-bound approach the number of integer variables determines the size of the search tree and influences the execution time of the algorithm. In our case, the number of integer variables is equal to the number of RBs and is limited by the available bandwidth (in LTE, the number is less than 100 for 20 MHz [29]). Thus, the optimization problem is solved in a reasonable time for practical scenarios.

C. User Association

In this section, we introduce the second building block of our coordinated framework. Exploiting the output of the spectrum allocation block, we compute the user peak rate and SINR per BS for each user in the network. Such computation takes into account the user radio conditions in terms of pathloss and interference, and is done according to the SINR expression (*i.e.,* (2) for LTE and (3) for mmWave) and peak rate (4), as introduced in Section II. Then, user association is performed following two widely adopted approaches, namely the *network centric* association and the *mobile-terminal centric* association. The network centric approach is implemented as a centralized optimization problem in Section III-C1, whereas the mobile-terminal centric approach is portrayed as a distributed non-cooperative game in Section III-C2.

Algorithm 1 Best Response Algorithm for femto cell RB allocation

Require: $\mathcal{J}_{femto}^{LTE}, \mathcal{K}, G_{jj'}, \forall (j,j') \in (\mathcal{J}_{femto}^{LTE})^2$
Require: Iteration $t \leftarrow 0$
1: $\boldsymbol{x}_j(0) \leftarrow \mathbf{0}, \forall j \in \mathcal{J}_{femto}^{LTE}$
2: **repeat**
3: $t \leftarrow t+1$
4: **for** $j \in \mathcal{J}_{femto}^{LTE}$ **do**
5: Solve $\mathcal{P}_j(\boldsymbol{x}_{-j}(t-1))$ in (10)
6: $\boldsymbol{x}_j(t) \leftarrow \boldsymbol{x}_j^*$ ▷ Optimal values of $\mathcal{P}_j(\boldsymbol{x}_{-j}(t-1))$
7: **end for**
8: **until** $(\boldsymbol{x}(t) == \boldsymbol{x}(t-1))$

Centralized approaches enable to compute a global maximum of the network utility that strikes a good compromise between efficiency and load balancing. Such solutions necessitate coordination and signaling between BSs. Distributed approaches enable autonomous users to maximize their own utility leading to complexity reduction at the cost of lower efficiency.

1) Centralized User Association: Building on the output of the spectrum allocation $\boldsymbol{x}$, the coordinated problem introduced in (7) boils down to the following centralized user association problem:

$$(\mathcal{P}(\boldsymbol{x})): \underset{\boldsymbol{\theta}}{\text{maximize}} \quad \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}^{LTE}} \theta_{ij}\log(\frac{\rho_{ij}(x_{jk})}{\sum_{i\in\mathcal{I}}\theta_{ij}}) + \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}^{mmW}} \theta_{ij}\log(\frac{\rho_{ij}}{\sum_{i\in\mathcal{I}}\theta_{ij}}) \tag{11a}$$

$$\text{subject to} \quad \sum_{j\in\mathcal{J}}\theta_{ij}\leq 1,\ \forall i\in\mathcal{I}, \tag{11b}$$

$$\theta_{ij}\geq 0,\ \forall i\in\mathcal{I}, \forall j\in\mathcal{J}. \tag{11c}$$

The objective (11a) is convex as a maximization of a concave function. Indeed, it can be written as:

$$\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}\theta_{ij}\log(\frac{\rho_{ij}}{\sum_{i\in\mathcal{I}}\theta_{ij}}) \tag{12a}$$

$$=\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}\theta_{ij}\log(\rho_{ij}) - \sum_{j\in\mathcal{J}}(\sum_{i\in\mathcal{I}}\theta_{ij})\log(\sum_{i\in\mathcal{I}}\theta_{ij}), \tag{12b}$$

where the first expression is linear (hence concave) and the second is concave as the opposite of the entropy function. Moreover, the constraints (11b) and (11c) are linear. Therefore, Problem $(\mathcal{P}(\boldsymbol{x}))$ is a convex optimization problem that can be solved very efficiently using solvers such as CVX [33]. However, such centralized solutions are efficient but highly computational. In fact, they require a central controller that collects information from BSs and UEs, optimizes parameters,

and sends signaling messages back to the BSs and UEs which can be cumbersome.

2) Distributed User Association: We propose to solve the distributed user association problem by having recourse to non-cooperative game theory. Non-cooperative game theory models the interactions between players competing for a common resource which is the set $\mathcal{J}$ of BSs. We define a multi-player game $\mathcal{G}^{UA}$ between the $|\mathcal{I}|$ users which are assumed to make their decisions without knowing the decisions of each other. Each user i strives to compute the amount of time θ_{ij} to be associated with BS j in a way to selfishly maximize its own utility. Note that, in this game, each user can associate with any BS, mmWave or LTE.

a) User Association Game: The formulation of the non-cooperative user association game $\mathcal{G}^{UA} = \langle \mathcal{I}, S^{UA}, \eta \rangle$ can be described as follows:

- The finite set of users $\mathcal{I} = \{1, \cdots, |\mathcal{I}|\}$.
- Each user $i \in \mathcal{I}$ strives to compute the amount of time θ_{ij} to be associated with BS $j, \forall j \in \mathcal{J}$. Hence, the action of user i is $\boldsymbol{\theta}_i = (\theta_{i1}, ..., \theta_{i|\mathcal{J}|})$. A *strategy profile* $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_{|\mathcal{I}|})$ specifies the strategies of all players.
- The space of strategies S^{UA} formed by the Cartesian product of each set of strategies $S^{UA} = S_1^{UA} \times S_2^{UA} \times ... \times S_{|\mathcal{I}|}^{UA}$, where the strategy space of any user i is $S_i^{UA} = \{0 \leq \theta_{ij} \leq 1, \ \forall j \in \mathcal{J} \text{ and } \sum_{j \in \mathcal{J}} \theta_{ij} \leq 1\}$.
- A set of utility functions

$$\eta(\boldsymbol{\theta}) = (\eta_1(\boldsymbol{\theta}), \eta_2(\boldsymbol{\theta}), \cdots, \eta_{|\mathcal{I}|}(\boldsymbol{\theta}))$$

that quantify users' utility for a given strategy profile $\boldsymbol{\theta}$. The utility function of any user i is given according to the objective function of the coordinated problem in (5):

$$\eta_i(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{-i}) = \sum_{j \in \mathcal{J}} \theta_{ij} \log(\frac{\rho_{ij}}{\theta_{ij} + \sum_{\substack{i' \in \mathcal{I} \\ i' \neq i}} \theta_{i'j}}). \quad (13)$$

For every user i, η_i is strictly concave w.r.t. $\theta_{ij}, \forall j \in \mathcal{J}$ and continuous w.r.t. $\theta_{lj}, l \neq i, \forall j \in \mathcal{J}$. Hence, a Pure NE exists and it is unique [34], which is a valuable asset for practical scenarios. In the following section, we will investigate how to reach this PNE.

b) Reaching Pure Nash Equilibriums: We prove hereafter that we are in presence of an exact potential game. This property enables us to use a greedy Best Response algorithm in order to reach the PNE according to [32].

Proposition 3.2: The game $\mathcal{G}^{UA}$ is an exact potential game.

Proof: The game $\mathcal{G}^{UA}$ is a continuous exact potential game as we have what follows:

$$\frac{\partial \eta_i}{\partial \theta_{ij}} = \frac{\partial V(\boldsymbol{\theta})}{\partial \theta_{ij}}, \forall i \in \mathcal{I}, \forall j \in \mathcal{J},$$

where the potential function V is given by:

$$V(\boldsymbol{\theta}) = \sum_{j \in \mathcal{J}} (\sum_{i \in \mathcal{I}} \theta_{ij} \log(\rho_{ij}) + \sum_{i \in \mathcal{I}} \theta_{ij} \log(\sum_{i \in \mathcal{I}} \theta_{ij})(|\mathcal{I}| - 2) - \sum_{i \in \mathcal{I}} (\sum_{\substack{u \in \mathcal{I} \\ u \neq i}} \theta_{uj}) \log(\sum_{\substack{u \in \mathcal{I} \\ u \neq i}} \theta_{uj})) \quad (14)$$

□

Exact potential games have a distinct computational advantage in that computing the PNE consists in maximizing the potential function. Therefore, the computation of an equilibrium is reduced to solving an optimization problem at each iteration of the Best Response dynamics, obviating the need for computational fixed point theory. Accordingly, at each iteration step t of the Best Response algorithm, user $i \in \mathcal{I}$ strives to find the following optimal association vector $\boldsymbol{\theta}_i^*(t)$ as a response to $\boldsymbol{\theta}_{-i}(t-1)$:

$$\boldsymbol{\theta}_i^*(t) = \underset{\boldsymbol{\theta}_i}{\operatorname{argmax}} \ \eta_i(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{-i}), \text{ subject to } \boldsymbol{\theta}_i \in S_i^{ua}, \quad (15)$$

which amounts to the following optimization problem:

$$(\mathcal{P}_i(\boldsymbol{\theta}_{-i})): \quad \max_{\boldsymbol{\theta}_i} \ \eta_i(\boldsymbol{\theta}) = \{\sum_{j \in \mathcal{J}} \theta_{ij} \log(\frac{\rho_{ij}}{\theta_{ij} + \sum_{\substack{i' \in \mathcal{I} \\ i' \neq i}} \theta_{i'j}})\} \quad (16a)$$

$$\text{subject to} \quad \sum_{j \in \mathcal{J}} \theta_{ij} \leq 1, \quad (16b)$$

$$\theta_{ij} \geq 0, \ \forall j \in \mathcal{J}. \quad (16c)$$

Problem $(\mathcal{P}_i(\boldsymbol{\theta}_{-i}))$ solved by each user $i \in \mathcal{I}$ at each iteration is a convex optimization problem. Indeed, the constraints (16b) and (16c) are linear, while the objective (16a) decomposes into two functions:

$$\eta_i = \sum_{j \in \mathcal{J}} \theta_{ij} \log(\rho_{ij}) - \sum_{j \in \mathcal{J}} \theta_{ij} \log(\theta_{ij} + \sum_{\substack{i' \in \mathcal{I} \\ i' \neq i}} \theta_{i'j}), \quad (17)$$

where the first expression is linear (hence concave) and the second is concave as the opposite of the entropy function. Convex optimization is a special class of mathematical problems that can be solved numerically very efficiently. In particular, we resort to the subgradient-based algorithm to solve Problem (16) in what follows.

c) Projected subgradient-based algorithm: In order to solve the convex problem $(\mathcal{P}_i(\boldsymbol{\theta}_{-i}))$ for each user $i \in \mathcal{I}$, we can have recourse to the subgradient method for constrained optimization. This algorithm takes the following form:

$$\theta_{ij}^{n+1} = \theta_{ij}^n + \delta^n g^n,$$

where n is the step number, δ^n a step size, and g^n a subgradient of the objective function in (16a). This subgradient is written as follows:

$$g^n = \nabla_{\theta_{ij}} \eta_i(\boldsymbol{\theta}_i^n, \boldsymbol{\theta}_{-i}^n) \quad (18a)$$

$$= \log(\frac{\rho_{ij}}{\theta_{ij}^n + \sum_{\substack{i' \in \mathcal{I} \\ i' \neq i}} \theta_{i'j}^n}) - \frac{\theta_{ij}^n}{\theta_{ij}^n + \sum_{\substack{i' \in \mathcal{I} \\ i' \neq i}} \theta_{i'j}^n}. \quad (18b)$$

The feasible set of user association vectors is a simplex defined by constraints (16b) and (16c). The projection of the subgradient on the simplex is straightforward and performed according to the algorithm in [35].

Algorithm 2 details the computation process for solving the user association problem in a distributed manner. Given

Algorithm 2 Best Response algorithm for user association

Require: $\mathcal{I}, \mathcal{J}, \mathcal{K}, \rho_{ij}, \forall (i,j) \in (\mathcal{I} \times \mathcal{J})$. Maximum tolerance for the subgradient algorithm $\epsilon \geq 0$ and for the Best Response algorithm $\epsilon' \geq 0$

Require: Iteration $t \leftarrow 0$

1: $\theta_{ij}(1) \leftarrow \frac{1}{|\mathcal{J}|}, \forall i \in \mathcal{I}, \forall j \in \mathcal{J}$
2: **repeat**
3: $t \leftarrow t+1$
4: **for** $i \in \mathcal{I}$ **do**
5: Step $n \leftarrow 0$
6: $\theta_{ij}^n(t) \leftarrow \theta_{ij}(t), \forall j \in \mathcal{J}$
7: **repeat** ▷ Solve $(\mathcal{P}_i(\boldsymbol{\theta}_{-i}))$ in (16)
8: **for** $j \in \mathcal{J}$ **do**
9: $\theta_{ij}^{n+1}(t) = \theta_{ij}^n(t) + \delta^n \left(\log(\frac{\rho_{ij}}{\theta_{ij}^n(t)+\sum_{\substack{i' \in \mathcal{I} \\ i' \neq i}} \theta_{i'j}(t)}) - \frac{\theta_{ij}^n(t)}{\theta_{ij}^n(t)+\sum_{\substack{i' \in \mathcal{I} \\ i' \neq i}} \theta_{i'j}(t)} \right)$
10: $\tilde{\theta}_{ij}^{n+1}(t) \leftarrow \text{Projection}(\theta_{ij}^{n+1}(t))$
11: **end for**
12: $n \leftarrow n+1$
13: **until** $|\tilde{\boldsymbol{\theta}}_i^{n+1}(t) - \tilde{\boldsymbol{\theta}}_i^n(t)| \leq \epsilon$
14: $\boldsymbol{\theta}_i(t+1) \leftarrow \tilde{\boldsymbol{\theta}}_i^{n+1}(t)$
15: **end for**
16: **until** $|\boldsymbol{\theta}(t+1) - \boldsymbol{\theta}(t)| \leq \epsilon'$

initial user association vectors, the Best Response algorithm iterates until convergence to a PNE (Line 16). Each iteration consists of computing the user association successively for all users in the network (Line 4). When dealing with user $i \in \mathcal{I}$, the optimal user association is computed using the projected subgradient method (Line 7-13). The subgradient algorithm converges when the variation in the user association vector between two successive steps is less than a predefined maximal tolerance ϵ (Line 13). The convergence of the gradient-based optimization is guaranteed and proved in 3.3. Similarly, the global user association problem converges when the variation in the UA vector between two successive iterations is less than a predefined maximal tolerance ϵ' (Line 16).

Proposition 3.3: The convergence of the subgradient method in the best response algorithm with constant step size is guaranteed.

Proof: According to [36], a constant step size is more convenient for distributed algorithms. In the latter case, the gradient algorithm converges to the optimal value provided that the step size is sufficiently small and that the objective function $\eta_i(\theta_i, \theta_{-i})$ has Lipschitz continuity property. The latter property is verified if the Hessian is bounded in the l_2 norm.

All Hessian diagonal values are equal to:

$$\lambda_{jj} = \frac{\partial^2 \eta_i}{\partial^2 \theta_{ij}} = -1 - \frac{d_{ij}}{\theta_{ij} + d_{ij}}, \tag{19}$$

where $d_{ij} = \sum_{\substack{i' \in \mathcal{I} \\ i' \neq i}} \theta_{i'j}$.

As non-diagonal values are null, $l_2 = \lambda_{jj}$. To establish the proof and show that l_2 norm is bounded, we need to distinguish two cases:

- Case 1: for a given user $i \in \mathcal{I}$, $\forall j \in \mathcal{J}$, there exists some $\theta_{i'j} \neq 0$ (i.e. at least one user is associated with BS j) and hence $0 < d_{ij} \leq |\mathcal{I}|$. Accordingly, $l_2 = \lambda_{jj}$ is bounded because $0 \leq \theta_{ij} \leq 1$:

$$-2 \leq \lambda_{jj} \leq \frac{-1 - 2 \times d_{ij}}{1 + d_{ij}} \tag{20}$$

- Case 2: for a given user $i \in \mathcal{I}$, $\exists j \in \mathcal{J}$ such as $\sum_{\substack{i' \in \mathcal{I} \\ i' \neq i}} \theta_{i'j} = 0$. Here, we need to separate again two subcases:
  1) $\theta_{ij} \neq 0$ which means that no user other than user i is associated with BS j. Accordingly, $\eta_i(\theta_i, \theta_{-i}) = \sum_{j \in \mathcal{J}} \theta_{ij} \log(\rho_{ij}) - \sum_{j \in \mathcal{J}} \theta_{ij} \log(\theta_{ij})$. Thus, $\frac{\partial^2 \eta_i}{\partial^2 \theta_{ij}} = \frac{-1}{\theta_{ij}}$ which means that the l_2 norm is bounded as $0 < \theta_{ij} \leq 1$.
  2) $\theta_{ij} = 0$ which means that there will be no traffic on BS j. Such a case is problematic because it will lead to an indeterminate expression in the subgradient of the objective function in (18b) $(0/0)$. However, we show that such a case will boil down to removing BS j from the gradient iteration of user i. In fact, we consider that at given iteration n, user i needs to optimize $\theta_{ij}^{(n)}$ and $\theta_{i'j}^{(n)} = 0$ for $i' \neq i \in \mathcal{I}$. If user i responds by setting $\theta_{ij}^{(n)}$ to zero in order to optimize its objective function in (17), θ_{ij} will remain null for further iterations. Indeed, if for the most suitable traffic condition on BS j (all concurrent users overlooked BS j), user i has decided not to select it, then it will avoid allocating traffic to that BS in the future when eventually other users might select it again, forcing the rate obtained by user i through BS j to further drop down.

□

In the Best Response algorithm 2, each user iteratively optimizes its utility using the projected-subgradient algorithm. The latter necessitates a set of parameters for initialization for each user $i \in \mathcal{I}$:

- The peak rates ρ_{ij} when associated to each BS $j \in \mathcal{J}$ (obtained as an output of the SA block).
- The UA vector $\boldsymbol{\theta}_i(t)$ (obtained locally at the previous iteration t).
- The expression $\sum_{\substack{i' \in \mathcal{I} \\ i' \neq i}} \theta_{i'j}(t)$.

The last expression can be written as $L_j(t) - \theta_{ij}(t)$, where $L_j(t) = \sum_{i \in \mathcal{I}} \theta_{ij}(t)$ is the load of BS $j \in \mathcal{J}$. This load information can be signaled on a broadcast channel by the BSs at each iteration t. As all the aforementioned parameters can be made available to the user, Algorithm 2 can be straightforwardly implemented in a distributed fashion.

IV. Performance Evaluation

In this section, we evaluate the impact of various network parameters, such as mmWave density, cell load, and user distribution and density. Finally, we offer valuable guidelines into practical 5G HetNet design.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2020.3022681, IEEE Transactions on Mobile Computing

A. Description of the Algorithms

We implement various algorithms for the coordinated Spectrum Allocation (SA) and User Association (UA) in 5G heterogeneous networks. These algorithms are presented hereafter:

1) BR-SA + BR-UA combines astute spectrum allocation as in Section III-B with distributed user association as in Section III-C2. This algorithm includes cell load estimation and applies distributed RB allocation to shield vulnerable femto cells from harmful interferences. Particularly, it implements Best Response dynamics for femto cell RB allocation (Algorithm 1). As for BR-UA, it implements a Best response algorithm with a projected-subgradient (Algorithm 2).
2) BR-SA + Cent-UA combines the same spectrum allocation as BR-SA + BR-UA with the centralized optimal user association as in III-C1. The optimal user association is based on convex optimization and solved with CVX [33].
3) BR-SA + Pow-UA combines the same spectrum allocation as previously with the state-of-the-art power-based user association. Precisely, each user is associated with the BS corresponding to the maximum received power.
4) CoCh-SA + PR-UA deploys the totality of the LTE spectrum in femto and macro cells with frequency reuse-1. It implements a peak rate user association where each user is associated with the BS delivering the highest peak rate given by (4).
5) SepCh-SA + SCFirst-UA equally shares the LTE spectrum between macro and femto tiers with frequency reuse-1. Users are associated with the BS delivering the maximum received power: a bias of 10% is added to the power received from small cells to artificially extend their coverage [37].
6) CoCh-SA + Pow-UA deploys the totality of spectrum in femto and macro cells with frequency reuse-1. User association is power based.

The choice of the aforementioned algorithms represents different strategies in spectrum allocation and user association. The first three algorithms challenge the savvy spectrum allocation made by BR-SA by combining it with different user association schemes: a selfish UA (BR-UA) that amounts to a global distributed solution, an optimal centralized user association (Cent-UA) that provides upper bound performances, and a basic state-of-the-art power-based association (Pow-UA). The last three algorithms combine static state-of-the-art spectrum allocation with basic UA schemes that tackle different aspects of network heterogeneity. PR-UA tackles heterogeneity in terms of spectrum size (20 MHz for LTE and 1 GHz for mmWave), SCFirst-UA tackles the heterogeneity in terms of cell size, and Pow-UA tackles the heterogeneity in terms of transmit power. In the following, we present a thorough assessment of the algorithms performance in the coordinated spectrum allocation and user association. We consider multiple criteria such as user rate distribution, traffic distribution, and global utility.

B. Simulation Context

We consider a three-tier wireless network comprising macro LTE cells, femto LTE cells, and mmWave cells over a square area of 1200 m x 1200 m shown in Fig. 2. We distribute seven tri-sectorized macro-BSs according to a hexagonal structure with an inter-BS distance of 500 m as in an urban environment. In each one of the 21 sectors, we uniformly allocate two geographical positions for femto and mmWave BSs.

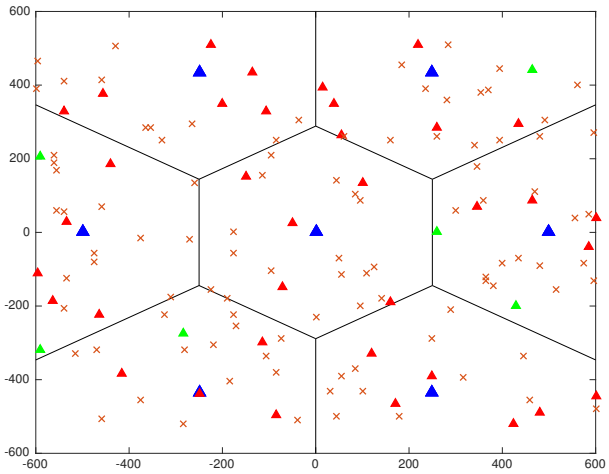

Fig. 2: Three-tier wireless network with LTE tri-sectorized macro BSs (in blue), LTE femto BSs (in red), and mmWave BSs (in green). Red crosses represent user positions.

In our simulations, the LTE system bandwidth equals 20 MHz with 100 RBs of 180 kHz. Multiple Input Multiple Output (MIMO) is supported to increase user rates eightfold through spatial multiplexing. When radio conditions are favorable, 8 different data streams are transmitted on 8 transmitting antennas and received on 8 receiving antennas, using the same RB. Moreover, channels are generated using the publicly-available MATLAB implementation of the WINNER Phase II Channel Model [38]. The shadow fading map follows a normal 2D space-correlated distribution, as in [39]: the Gaussian random variable has a zero mean and a standard deviation of 10 dB.

As for the mmWave system, we consider a bandwidth of 1 GHz with a central frequency of 73 GHz. Channels are generated using the model introduced by [40] following the outdoor propagation measurement campaign in New York city. We randomly consider 10% of the links between users and mmWave BSs to have line of sight propagation. Here also, the shadow fading map follows a normal distribution: the Gaussian random variable has a zero mean and a standard deviation of 10 dB.

In the heterogeneous network, the transmit power of the deployed antennas equals 42 dBm for macro BSs, 23 dBm for femto BSs, and 30 dBm for mmWave. The transmit and receive antenna gains equal 15 dBi and 0 dBi, respectively for all types of devices.

The following numerical results are obtained for 20 runs of each algorithm. In order to compare these results with statistical distributions, we make use of boxplots: on each

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2020.3022681, IEEE Transactions on Mobile Computing

boxplot, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points which are not considered outliers, and outliers are plotted individually. We also make use of cumulative distribution functions computed based on the empirical measure of the total runs of the algorithms.

C. Reference Scenario

We start by defining a reference scenario where users are generated according to a random uniform distribution. In this scenario, the per-cell cluster C_j of BS j is the set of BSs $j' \in C_j$ where $\gamma_{jj'} \geq \gamma_{threshold}$ with $\gamma_{threshold} = -110$ dB. Further, 15% of the generated small-cell positions are dedicated to mmWave BSs (according to a uniform random distribution), the remaining for femto BSs. In this section, we perform an exhaustive comparison of the various algorithms using the settings of the reference scenario. Then, this scenario is declined with multiple variations to stress out the impact of different factors such as the mmWave BS density, the user distribution and density, and finally, the femto cluster size.

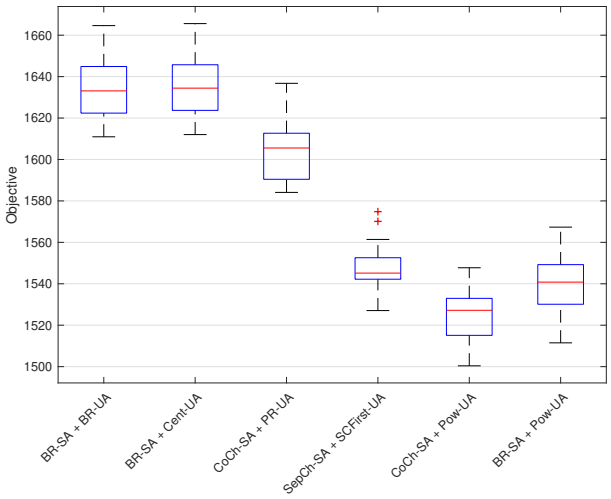

Fig. 3: Objective Value for reference scenario

We start by analyzing the performances of the simulated algorithms in Fig. 3 and Fig. 4. Figure 3 displays the numerical values of the objective in (7a) for the coordinated problem, whereas Fig. 4 shows the percentage of users associated with macro, femto, and mmWave BSs. This percentage is computed for macro BSs as follows

$$(\sum_{j \in \mathcal{J}_{macro}^{LTE}} \sum_{i \in \mathcal{I}} \theta_{ij}) / (\sum_{j \in \mathcal{J}} n_j).$$

The percentage of users associated with femto and mmWave BSs are derived similarly.

In Fig. 3, we note that BR-SA + BR-UA and BR-SA + Cent-UA achieve the best performances. First, BR-SA benefits from the devised cell load estimation and interference minimization. This increases the peak rate and consequently the value of the objective function. Using an optimal centralized user association in conjunction with BR-SA (as in BR-SA + Cent-UA) leads to the best performance owing to the high peak rates

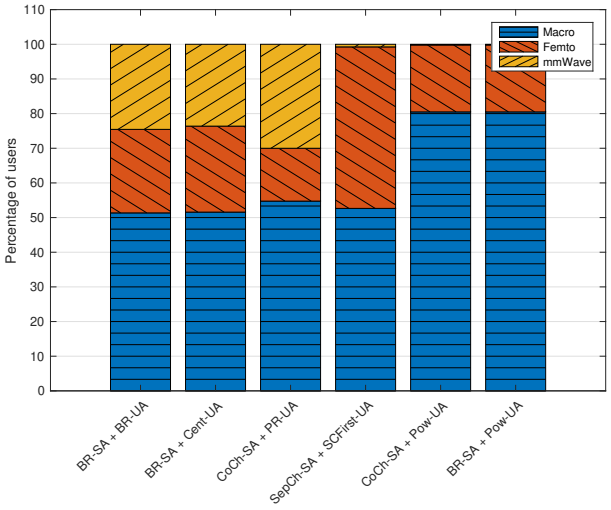

Fig. 4: User association percentage for reference scenario

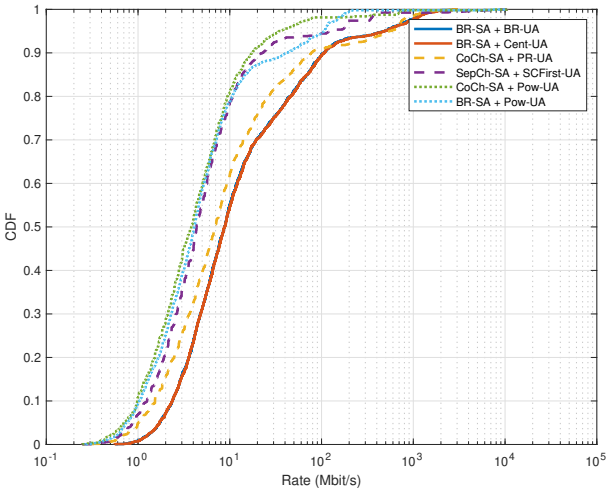

Fig. 5: User rate distribution for reference scenario

made available by BR-SA. Moreover, the distributed BR-UA matches the optimal Cent-UA and the algorithm BR-SA + BR-UA achieves equivalent performances compared to BR-SA + Cent-UA. In Fig. 4, BR-SA + BR-UA and BR-SA + Cent-UA depict equivalent user association percentages where users are equally shared among macro cells and small cells.

Simulation results show the limitations of the widespread state-of-the-art power based association. In Fig. 3, the objective value obtained by BR-SA + Pow-UA is low despite the high efficiency of the spectrum allocation performed with BR-SA. In such a heterogeneous network, the power based user association Pow-UA incites users to select the BS with the highest transmit power thus leading to congestion on the LTE macro BSs. This is confirmed by Fig. 4 where 80% of users are associated with macro BSs and almost no users select mmWave.

The limitation of power based association is exacerbated when CoCh-SA is used to statically allocate the spectrum in CoCh-SA + Pow-UA. In this case, reuse-1 spectrum allocation is agnostic to cell load and users associated to LTE BSs suffer from magnified mutual interference between macro and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2020.3022681, IEEE Transactions on Mobile Computing

femto. Such interference lowers the peak rates available in the network and consequently the objective value after user association as shown in Fig. 3.

Interestingly, CoCh-SA performs better in CoCh-SA + PR-UA than in CoCh-SA + Pow-UA in terms of objective value as seen in Fig. 3. The former associates users with the BS delivering the highest peak rate taking full benefit from technologies such as mmWave, having short transmission range but large bandwidth. In fact, 30% of users are associated with mmWave BSs in CoCh-SA + PR-UA as in Fig. 4.

Finally, SepCh-SA + SCFirst-UA achieves a lower objective than CoCh-SA + PR-UA despite the bias towards small cells and the interference cancellation obtained by a separate spectrum allocation for macro and femto BSs. First, the spectrum separation between macro and femto BSs is independent of the user distribution. Moreover, the reuse-1 strategy does not protect femto users from harmful co-tier interference. Second, the *small-cell first* user association does not lead to the suitable user offload in a heterogeneous setting as seen in Fig. 4.

In Fig. 5, we plot the cumulative distribution function of user rates for the different algorithms. The median rate for BR-SA + BR-UA and BR-SA + Cent-UA is equal to 9 Mbit/s, followed by 7 Mbit/s for CoCh-SA + PR-UA, and only 3 Mbit/s for the remaining algorithms. The plot shows that around 50% of users have more than 10 Mbit/s for BR-SA + BR-UA and BR-SA + Cent-UA, 40% for CoCh-SA + PR-UA, and less than 20% for the other algorithms. The advantage of the best performing algorithms increases between 20 Mbit/s and 100 Mbit/s. For instance, 20% of users obtain more than 50 Mbit/s with these two algorithms, 15% with CoCh-SA + PR-UA, and only 1% with CoCh-SA + Pow-UA. Particularly, BR-SA + Pow-UA outperforms the state-of-the-art algorithms around 50 Mbit/s owing to the astute spectrum allocation that reduces the harmful interference among LTE BSs.

D. Impact of mmWave Density

In this section, we assess the impact of mmWave BSs density. We consider two scenarios where 5% and 25% of the generated small-cell positions are dedicated to mmWave BSs (according to a uniform random distribution), the remaining for femto BSs, and compare them with the reference scenario (15% for mmWave BSs).

Except the CoCh-SA + Pow-UA algorithm that is insensible to the spectrum opportunity brought by mmWave, the other algorithms have seen their objective largely increased in proportion to the mmWave density as shown in Fig. 6. This behavior is confirmed through Fig. 7 where we display the percentage of users associated to the various BSs. In the CoCh-SA + Pow-UA algorithm, the mmWave technology fails to attract users, contrary to the CoCh-SA + PR-UA algorithm where mmWave BSs absorb almost half of the traffic for the highest density. The BR-SA + BR-UA algorithm is more balanced as it shelters femto BSs from deleterious interference via intelligent spectrum allocation. This enables femto BSs to remain appealing for users: around 25% of users select femto BSs in the reference model for BR-SA + BR-UA against 15% for CoCh-SA + PR-UA. We note that this percentage is further

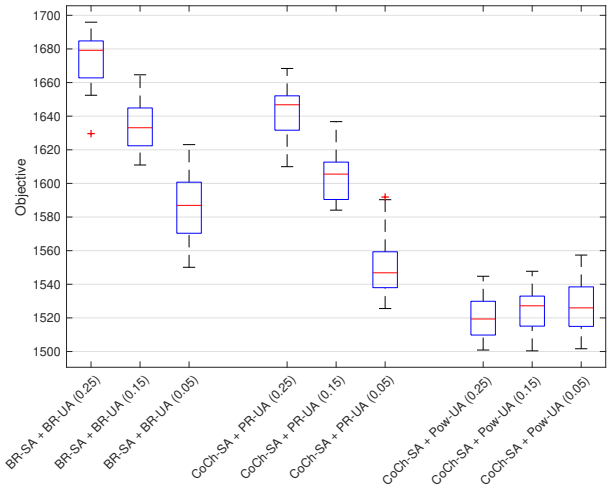

Fig. 6: Value of the objective for different mmWave densities

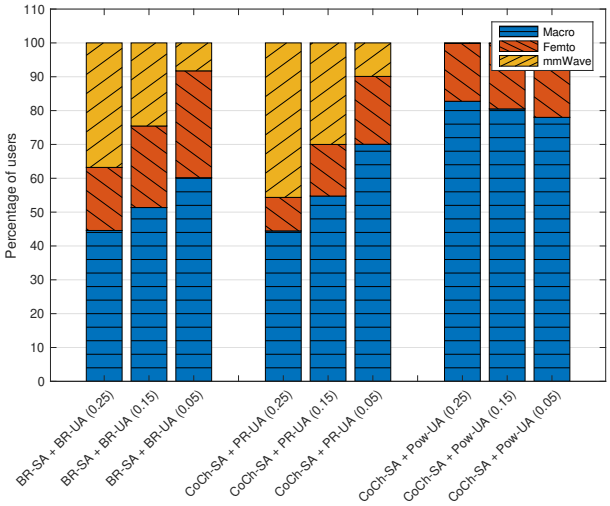

Fig. 7: User association percentage for different mmWave densities

reduced when the mmWave density increases: only 10% for CoCh-SA + PR-UA against 20% for BR-SA + BR-UA.

The user rate represented in Fig. 8 shows the invariance of the state-of-the-art algorithm CoCh-SA + Pow-UA, whereas the median rate increases significantly with the mmWave density for the two other algorithms. Precisely, CoCh-SA + PR-UA increases its median rate by 5 Mbit/s from low to high mmWave density, while BR-SA + BR-UA records an increase of 8 Mbit/s. We note that the intelligent coordinated algorithm BR-SA + BR-UA enables to take full advantage of the heterogeneous deployment. For a low mmWave density, the third quartile is equal to 10 Mbit/s for CoCh-SA + PR-UA and 12 Mbit/s for BR-SA + BR-UA. For a high mmWave density, the performance gap increases significantly: the third quartile is equal to 30 Mbit/s for CoCh-SA + PR-UA and 45 Mbit/s for BR-SA + BR-UA.

E. Impact of the User Distribution

In this section, we generate users according to a Gaussian distribution with a standard deviation of 45 centered at the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2020.3022681, IEEE Transactions on Mobile Computing

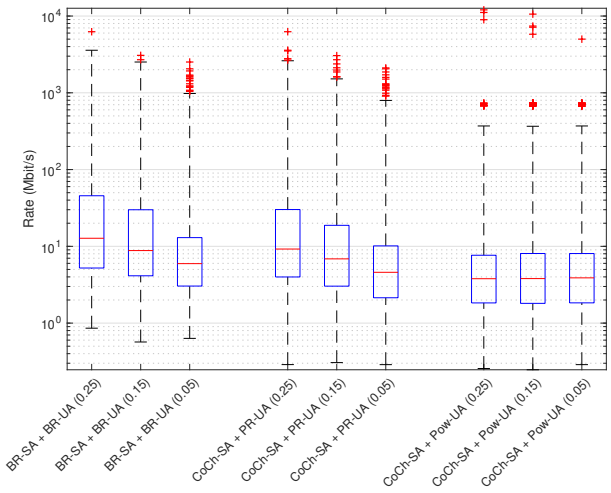

Fig. 8: User rate distribution for different mmWave densities

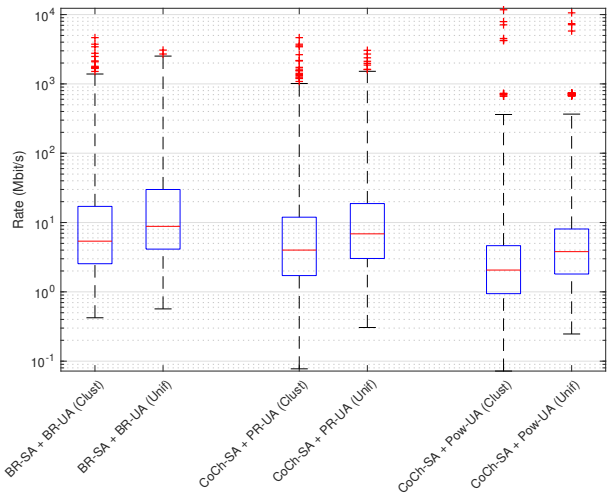

Fig. 10: User rate distribution for different user distributions

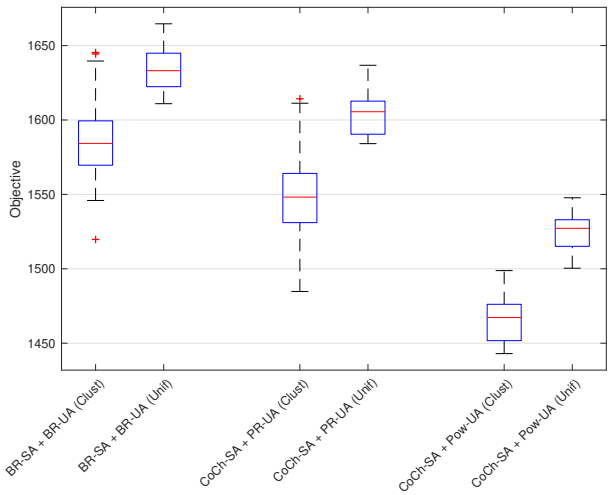

Fig. 9: Value of the objective for different user distributions

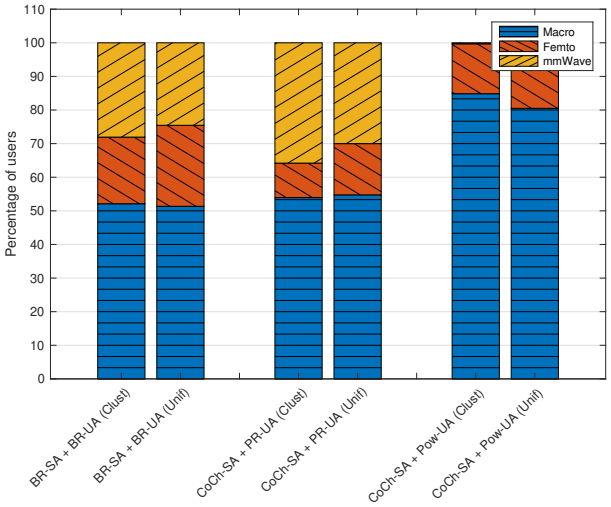

Fig. 11: User association percentage for different user distributions

origin of the network represented in Fig. 2. This scenario represents a crowded space where the user density is very high near the central macro BS. In Fig. 9, we compare the objective value as computed by the selected algorithms for the reference scenario (uniform distribution of users) and the crowd scenario (Gaussian distribution of users). We note a performance deterioration in the crowd scenario: some BSs in the network become congested while others are underutilized. This is exacerbated for the CoCh-SA + Pow-UA algorithm where the power based UA restricts the number of candidate BSs. This is further highlighted in Fig. 10 where the median rate of BR-SA + BR-UA decreases from around 9 Mbit/s to 5 Mbit/s and the median rate of CoCh-SA + Pow-UA drops from around 4 Mbit/s to 2 Mbit/s. Finally, in Fig. 11, we note that in the crowd scenario, the percentage of users associated with mmWave BSs becomes slightly higher (except for power based user association). In fact, the number of users that are in the mmWave coverage increases for the crowded region.

F. Impact of Femto Cluster Size

We assess in this section the impact of femto cluster size on performances. In particular, we show that through cluster size, we can control spectrum reuse.

a) Fine-tuning of Cluster Size: In this scenario, we consider different values for the $\gamma_{threshold}$ that defines the cluster size relative to each femto BS: -110 dB for the reference scenario, -80 dB for smaller clusters, and -120 dB for larger ones. To assess the impact of this threshold, we display in Fig. 12 the cumulative distribution function for the user rates.

We note a low discrepancy between the different cases except for rates ranging between 2 and 30 Mbit/s. Using a higher threshold in comparison with the reference scenario shrinks the cluster size. This increases the amount of available resource blocks per femto BS, at the expense of magnified interference. Precisely, the high interference level leads to fewer users having rates between 5 and 30 Mbits/s compared with the reference scenario. Contrarily, when decreasing the threshold to -120 dB, the cluster size is increased, thus

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2020.3022681, IEEE Transactions on Mobile Computing

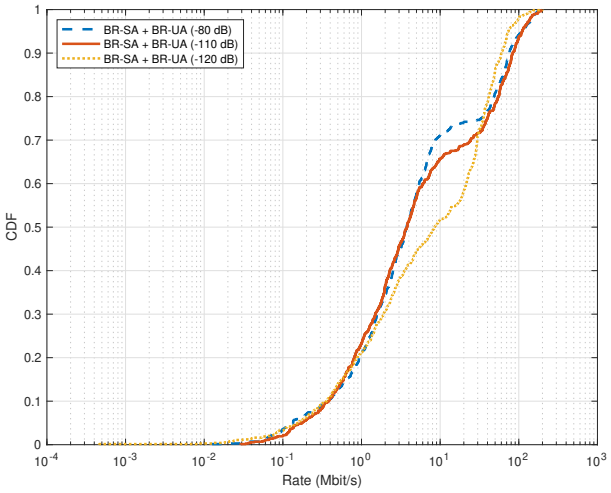

Fig. 12: Impact of reuse path loss threshold

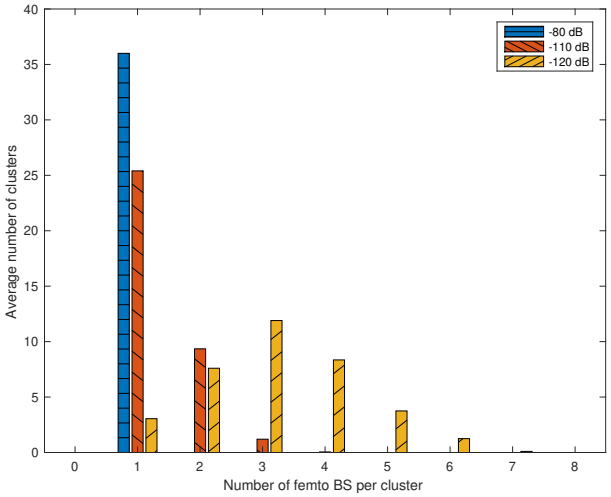

Fig. 13: Impact of reuse path loss threshold on cluster size

reducing the number of available resources per femto BS. This leads to fewer users having rates above 30 Mbit/s. However, as the interference level is lowered, more users exploit the MIMO spatial multiplexing gain, and consequently more users have rates above 2 till 30 Mbit/s. Thus, the threshold is a powerful tool to balance between interference and spectrum reuse. A large threshold reduces the size of each per-cell cluster enabling higher spectrum reuse at the cost of higher co-tier interference and lower spectral efficiency, and vice versa.

b) Cluster Size and Femto Demand: Figures 13 and 14 respectively illustrate the impact of $\gamma_{threshold}$ on cluster size and femto demand. As discussed earlier, a larger threshold value reduces the cluster size and increases the number of resource blocks per femto BS. We note, in Fig. 14, the very high number of femto BSs with zero demand. These BSs may be turned off, so as to reduce network power consumption while still achieving the best performances. In fact, although these BSs serve no users, the objective function value of BR-SA + BR-UA significantly exceeds that of CoCh-SA + PR-UA, SepCh-SA + SCFirst-UA, and CoCh-SA + Pow-UA, where all femto BSs are provided RBs (cf. Fig. 3). Moreover, numerical results show that providing one RB to the femto BS with zero demand yields a decrease in the spectral efficiency and in the objective function value. This confirms the efficiency of our femto cell load estimation.

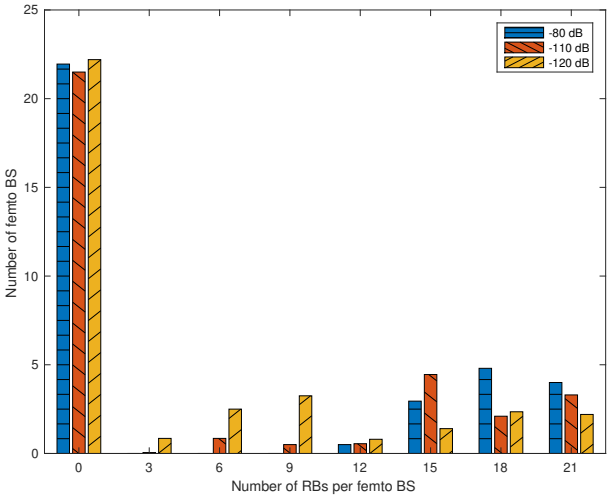

Fig. 14: Impact of reuse path loss threshold on femto demand

G. Complexity Analysis

Let us examine the complexity of the algorithms proposed in this paper. Our coordinated framework operates in two steps: a spectrum allocation followed by a user association.

In all heuristic user association algorithms (Pow-UA, PR-UA, SCFirst-UA), every user computes a given metric per BS and compare the metrics provided by all BSs to choose its allocation. Thus, the complexity of these algorithms is in $O(|\mathcal{I}| \cdot |\mathcal{J}|)$. The game theory based approach BR-UA proceeds in iterations. Given initial user association vectors, the Best Response algorithm iterates until convergence to a NE. Each iteration consists of computing the user association successively for all users in the network. Thus, it is important

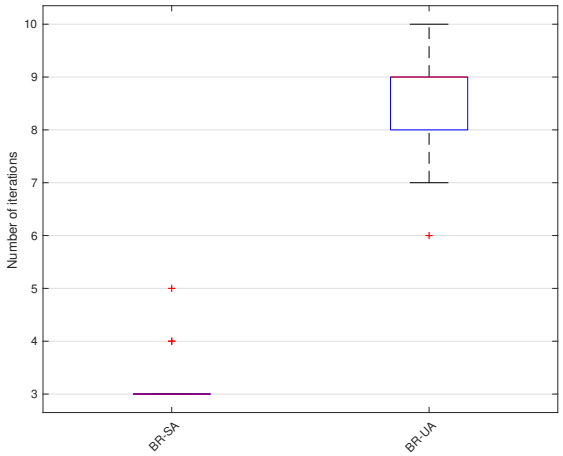

Fig. 15: Convergence rate of BR algorithm

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2020.3022681, IEEE Transactions on Mobile Computing

to evaluate the number of iterations in order to assess the complexity of such algorithm. For this, we have performed exhaustive simulations and plotted the distribution of the number of iterations in Fig. 15. We note that the first quartile of the iterations is equal to 8, the median is equal to 9 (corresponding to the third quartile), and the values never exceed 10. This ensures in practice a fast convergence of our algorithm.

Similar to the spectrum allocation, the game theory based approach, BR-SA, proceeds in iterations. At each iteration, each femto cell minimizes its cost function in response to the strategies of other femto cells in the previous iteration. Fig. 15 shows that the median number of iterations is equal to 3. This median value is also equal to the first and third quartiles.

H. Guidelines

For clarity, we sum up in this section the major contributions of our coordinated framework for SA and UA in novel 5G HetNets. Through this work, we proved that our devised framework yields load-aware RRM that prevents over-dimensioning radio resources, or defectively associating users to crowded cells.

In our exhaustive simulations, the Best Response algorithm for spectrum allocation enabled mitigating the impact of harmful interference on femto BSs. When coupled with our pertinent load estimation method as in BR-SA, we attained the best performances in terms of global objective and user rates. In fact, the load estimation succeeded in tailoring the spectrum allocation to the user distribution. As for the user association, the distributed approach in BR-UA matched perfectly the optimal centralized Cent-UA. Consequently, the BR-SA + BR-UA algorithm achieved the highest performance compared to all other algorithms.

In the heterogeneous wireless network, we showed the prevalence of coordinated algorithms that capture the diverse characteristics of the radio propagation in terms of radio channel model, transmit power, and bandwidth. In particular, we shed light on the deficiencies of power based user association Pow-UA, and the limitation of peak rate based association PR-UA that both failed to consider the totality of the aforementioned characteristics.

When dealing with the widely used small-cell first user association SCFirst-UA, we noted its drawbacks as highlighted in SepCh-SA + SCFirst-UA. In fact, the bias towards small cells did not enable to benefit from resources available in mmWave BSs. On the contrary, it provoked congestion on the overloaded femto BSs.

Considering the global performance, CoCh-SA + PR-UA showed the closest results compared with the best coordinated algorithms (*i.e.,* BR-SA + BR-UA). However, when the density of mmWave BSs is high, their appeal increased at the expense of femto BSs. This provoked traffic imbalance and caps the performance of CoCh-SA + PR-UA.

Finally, in crowd scenarios, the global objective of the network and the median rates diminished for all algorithms compared to a uniform user distribution. This congestion impact is exacerbated for static spectrum allocation (without load estimation) and power based user association (agnostic to load) as with the CoCh-SA + Pow-UA algorithm.

V. Massive MIMO in mmWave Cells

Massive MIMO (mMIMO) [41] and mmWave are regarded as two enabling 5G technologies. In this section, these two technologies are combined to improve capacity. We assume that mMIMO mmWave BS $j \in \mathcal{J}^{mmW}$ employs time-division duplexing (TDD), has M_j antennas, and simultaneously communicates with B_j single-antenna users over the same time-frequency resource ($M_j \gg B_j \geq 1$). In addition, relying on channel reciprocity in TDD cells, the downlink channel is estimated based on uplink pilots, whose number is equal to B_j.

Further, mmWave BS j is considered to use linear zero-forcing beamforming to transmit B_j user streams with equal power assignment. Consequently, the SINR observed by user $i \in \mathcal{I}$, when connected to mmWave BS $j \in \mathcal{J}^{mmW}$, can be rewritten as:

$$SINR_{ij}^{mmW} = \frac{M_j - B_j + 1}{B_j} \frac{p_j g_j g_i \gamma_{ij}}{\sum_{\substack{j' \in \mathcal{J}^{mmW} \\ j' \neq j}} p_{j'} g_{j'} g_i \gamma_{ij'} + \tilde{p}_N}, \tag{21}$$

Moreover, the downlink achievable rate ρ_{ij} of user $i \in \mathcal{I}$, when associated with BS $j \in \mathcal{J}^{mmW}$, is expressed as follows:

$$\rho_{ij} = \psi W_j \log_2(1 + SINR_{ij}^{mmW}), \tag{22}$$

where ψ is between 0 and 1 and represents the fraction of resources used for downlink data transmission. The remaining fraction is used either for uplink data transmission or channel estimation.

The coordinated spectrum allocation and user association problem is thus reformulated as:

$$\begin{aligned}
\underset{\boldsymbol{\theta}, \boldsymbol{x}}{\text{maximize}} \quad & \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}^{LTE}} \theta_{ij} \log\left(\frac{\rho_{ij}(x_{jk})}{\sum_{i \in \mathcal{I}} \theta_{ij}}\right) \\
& + \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}^{mmW}} \theta_{ij} \log(\rho_{ij}) && \text{(23a)} \\
\text{subject to} \quad & \sum_{j \in \mathcal{J}} \theta_{ij} \leq 1, \ \forall i \in \mathcal{I}, && \text{(23b)} \\
& \sum_{i \in \mathcal{I}} \theta_{ij} \leq B_j, \ \forall j \in \mathcal{J}^{mmW}, && \text{(23c)} \\
& \theta_{ij} \geq 0, \ \forall i \in \mathcal{I}, \forall j \in \mathcal{J}, && \text{(23d)} \\
& x_{jk} \in \{0, 1\}, \ \forall j \in \mathcal{J}^{LTE}, \forall k \in \mathcal{K}. && \text{(23e)}
\end{aligned}$$

In the new objective function (23a), we note the absence of scheduling in the part that covers the mmWave BSs $\sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}^{mmW}} \theta_{ij} \log(\rho_{ij})$, owing to the mMIMO technology. In fact, users associated to mmWave BS j are served simulatenously as long as $\sum_{i \in \mathcal{I}} \theta_{ij} \leq B_j$, as stated in constraints (23c). The optimization problem formulated in (23) is still a Mixed Integer Non-Linear Program (MINLP).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2020.3022681, IEEE Transactions on Mobile Computing

Building on the output of the spectrum allocation $\boldsymbol{x}$, the coordinated problem in (23) boils down to the following centralized user association problem:

$$(\mathcal{P}'(\boldsymbol{x})) : \underset{\boldsymbol{\theta}}{\text{maximize}} \quad \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}^{LTE}} \theta_{ij}\log\left(\frac{\rho_{ij}(x_{jk})}{\sum_{i\in\mathcal{I}}\theta_{ij}}\right) + \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}^{mmW}} \theta_{ij}\log(\rho_{ij}) \tag{24a}$$

$$\text{subject to} \quad \sum_{j\in\mathcal{J}} \theta_{ij} \leq 1,\ \forall i \in \mathcal{I}, \tag{24b}$$

$$\sum_{i\in\mathcal{I}} \theta_{ij} \leq B_j,\ \forall j \in \mathcal{J}^{mmW}, \tag{24c}$$

$$\theta_{ij} \geq 0,\ \forall i \in \mathcal{I}, \forall j \in \mathcal{J}. \tag{24d}$$

The objective (24a) is a concave function. Indeed, the part that covers LTE is concave similarly to (11a), whereas the part that concerns mmWave $\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}^{mmW}} \theta_{ij}\log(\rho_{ij})$ is linear in $\theta_{ij}, \forall j \in \mathcal{J}^{mmW}$ and hence concave. Moreover, all constraints are linear. Therefore, Problem $(\mathcal{P}'(\boldsymbol{x}))$ is a convex optimization problem.

In what follows, we assess the impact of massive MIMO combined with mmWave. For illustration, we consider the BR-SA + Cent-UA algorithm, as it achieves the best performances. We also assume that mmWave BS $j \in \mathcal{J}^{mmW}$ has the following parameters: $M_j = 100$, $B_j = 20$, and $W_j = 1$ GHz. We examine two scenarios, that differ in the fraction of resources used for downlink data transmission, and compare them with the reference scenario (without mMIMO). In the first one, ψ equals 0.45: radio resources are alternately shared between downlink data transmission and uplink data transmission. An additional fraction is further sacrificed for channel estimation. In the second scenario, downlink and uplink traffics are assumed to be extremely asymmetric, and all resources are ideally dedicated to downlink data transmission. Thus, ψ equals 1, maximizing the cell downlink capacity.

Since mMIMO boosts spectral efficiency and capacity, mMIMO mmWave BSs attract more users, offloading LTE cells, as shown in Fig. 16. With mMIMO, around 40% of users select mmWave BSs against an average of 25% for the case without mMIMO. Interestingly, when all resources are dedicated to downlink data transmission ($\psi = 1$), mmWave BSs fail to attract more users in comparison with the first scenario ($\psi = 0.45$). This highlights the coverage limitation of mmWave frequencies.

Moreover, higher median objective function values are observed with mMIMO in Fig. 17. As more users benefit from mMIMO and mmWave technologies, and less users compete for LTE resources, the objective value increases. In the first scenario, mmWave BSs use less than half their resources for downlink transmission, in comparison with the reference scenario. However, through improving spectral efficiency, the massive MIMO technology compensates for the lack of radio resources and yields slightly higher median objective value. Consequently, with mMIMO, similar performances can be achieved at lower costs. Further, in the second scenario, all mmWave resources are dedicated to downlink transmission.

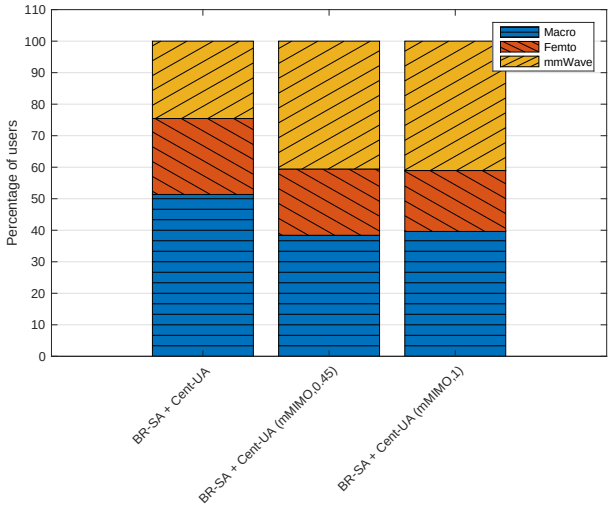

Fig. 16: Impact of mMIMO on user association percentage

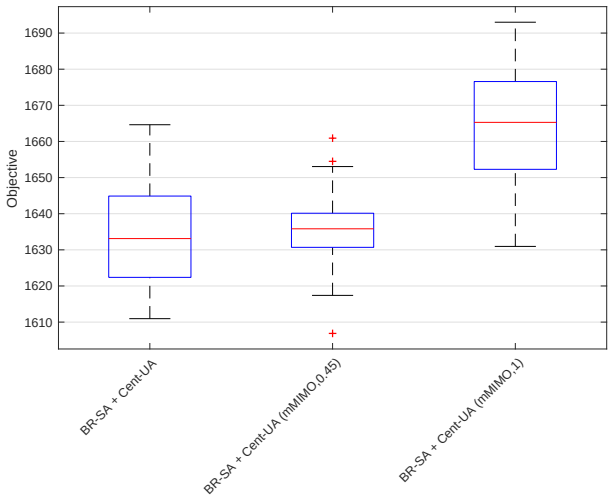

Fig. 17: Impact of mMIMO on the objective value

Users associated with mmWave BSs take full benefit of mMIMO and mmWave resources, contributing to a significant increase in the objective value.

Furthermore, as mMIMO improves mmWave spectral efficiency and capacity and reduces LTE load, median user rates increase in comparison with the reference scenario (cf. Fig. 18): 9 Mbit/s for the reference scenario, 9.5 Mbit/s for the first mMIMO scenario, and 12.5 Mbit/s for the second mMIMO scenario. Besides, the lowest user rate value improves to around 0.85 Mbit/s for the first mMIMO scenario and around 1 Mbit/s for the second mMIMO scenario. To conclude, using mMIMO allows our BR-SA + Cent-UA algorithm to achieve even higher performances in 5G HetNets (cf. second mMIMO scenario), or relatively similar performances although with fewer mmWave resources (cf. first mMIMO scenario).

VI. Conclusion

In this work, we tackled the problem of spectrum allocation and user association in 5G heterogeneous networks consisting of macro cells and small cells deployment with both mmWave

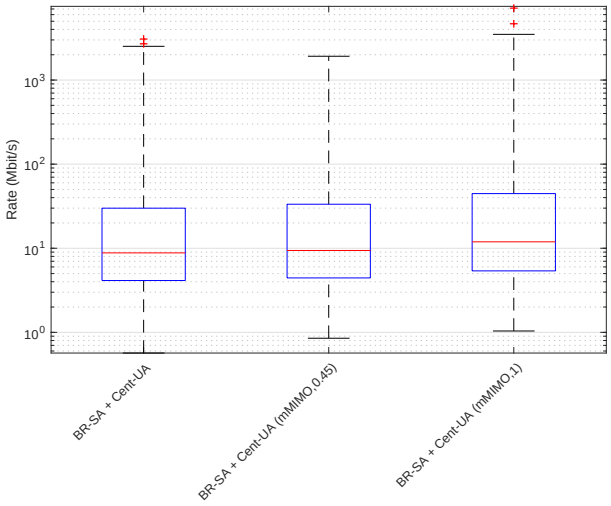

Fig. 18: Impact of mMIMO on user rate

and traditional sub-6GHz technology such as LTE. We introduced a coordinated framework that enabled tailoring the spectrum allocation to the user distribution and performing user association accordingly. In particular, we devised an original method for estimating cell load. This method served as a guideline to allocate the spectrum in a way to account for the user distribution and strike a good balance between spectral efficiency and spectrum reuse. The spectrum allocation for LTE small cells was formulated as a non-cooperative game. We proved the existence of pure Nash equilibrium that are attained by Best Response dynamics. The user association was addressed following two approaches: a centralized approach solved using convex optimization tools, and a distributed approach portrayed as a non-cooperative game. The unique Nash equilibrium of the latter game is attained by a fast converging Best Response algorithm. We implemented in our coordinated framework various spectrum allocation and user association algorithms following centralized, distributed, or basic state-of-the art approaches. Through extensive simulation, we assessed the performance of the devised algorithms considering relevant criteria such as BSs density, user density, traffic distribution, cluster size, and global objective. We further succeeded in offering valuable guidelines into 5G HetNet design. Finally, we assessed the impact of massive MIMO combined with mmWave in typical 5G HetNets.

REFERENCES

[1] S. Andreev, V. Petrov, M. Dohler, and H. Yanikomeroglu, "Future of ultra-dense networks beyond 5g: Harnessing heterogeneous moving cells," *IEEE Communications Magazine*, vol. 57, no. 6, pp. 86–92, June 2019.

[2] M. Cheng, J. Wang, Y. Wu, X. Xia, K. Wong, and M. Lin, "Coverage analysis for millimeter wave cellular networks with imperfect beam alignment," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 9, pp. 8302–8314, Sep. 2018.

[3] D. Liu, L. Wang, Y. Chen, M. Elkashlan, K. Wong, R. Schober, and L. Hanzo, "User association in 5g networks: A survey and an outlook," *IEEE Communications Surveys Tutorials*, vol. 18, no. 2, pp. 1018–1044, 2016.

[4] R. Madan, J. Borran, A. Sampath, N. Bhushan, A. Khandekar, and T. Ji, "Cell association and interference coordination in heterogeneous lte-a cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 9, pp. 1479–1489, December 2010.

[5] I. Guvenc, "Capacity and fairness analysis of heterogeneous networks with range expansion and interference coordination," *IEEE Communications Letters*, vol. 15, no. 10, pp. 1084–1087, October 2011.

[6] I. Guvenc, M. Jeong, I. Demirdogen, B. Kecicioglu, and F. Watanabe, "Range expansion and inter-cell interference coordination (icic) for picocell networks," in *2011 IEEE Vehicular Technology Conference (VTC Fall)*, Sep. 2011.

[7] S. Singh and J. G. Andrews, "Joint resource partitioning and offloading in heterogeneous cellular networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 2, pp. 888–901, February 2014.

[8] Y. Lin, W. Bao, W. Yu, and B. Liang, "Optimizing user association and spectrum allocation in hetnets: A utility perspective," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 6, pp. 1025–1039, June 2015.

[9] S. Deb, P. Monogioudis, J. Miernik, and J. P. Seymour, "Algorithms for enhanced inter-cell interference coordination (eicic) in lte hetnets," *IEEE/ACM Transactions on Networking*, vol. 22, no. 1, pp. 137–150, Feb 2014.

[10] D. Fooladivanda and C. Rosenberg, "Joint resource allocation and user association for heterogeneous wireless cellular networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 1, pp. 248–257, January 2013.

[11] Q. Kuang, W. Utschick, and A. Dotzler, "Optimal joint user association and multi-pattern resource allocation in heterogeneous networks," *IEEE Transactions on Signal Processing*, vol. 64, no. 13, pp. 3388–3401, July 2016.

[12] W. Bao and B. Liang, "Structured spectrum allocation and user association in heterogeneous cellular networks," in *IEEE INFOCOM 2014 - IEEE Conference on Computer Communications*, April 2014, pp. 1069–1077.

[13] Y. Li, M. Sheng, Y. Sun, and Y. Shi, "Joint optimization of bs operation, user association, subcarrier assignment, and power allocation for energy-efficient hetnets," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 12, pp. 3339–3353, Dec 2016.

[14] B. Zhuang, D. Guo, E. Wei, and M. L. Honig, "Scalable spectrum allocation and user association in networks with many small cells," *IEEE Transactions on Communications*, vol. 65, no. 7, pp. 2931–2942, July 2017.

[15] H. Y. Kim, H. Kim, Y. H. Cho, and S. Lee, "Self-organizing spectrum breathing and user association for load balancing in wireless networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 5, pp. 3409–3421, May 2016.

[16] H. Zhang, N. Prasad, and S. Rangarajan, "Method and systems for dynamic and configuration based fractional frequency reuse for uneven load distributions," March 2011, US 2001/0070911 A1.

[17] 3GPP TR 38.913, "Scenarios and Requirements for Next Generation Access Technologies, V0.2.0," February 2016.

[18] K. Sakaguchi, G. K. Tran, H. Shimodaira, S. Nanba, T. Sakurai, K. Takinami, I. Siaud, E. C. Strinati, A. Capone, I. Karls, R. Arefi, and T. Haustein, "Millimeter-wave evolution for 5g cellular networks," *IEICE Transactions*, vol. 98-B, no. 3, pp. 388–402, 2015.

[19] M. Ando, T. Taniguchi, M. Noda, A. Yamaguchi, K. Sakaguchi, and G. K. Tran, "Poc of mmwave (40 and 60 ghz) integrated 5g heterogeneous networks," in *European Wireless 2016; 22th European Wireless Conference*, May 2016.

[20] S. Okasaka, R. J. Weiler, W. Keusgen, A. Pudeyev, A. Maltsev, I. Karls, and K. Sakaguchi, "Proof-of-concept of a millimeter-wave integrated heterogeneous network for 5g cellular," *Sensors*, vol. 16, no. 9, August 2016.

[21] Y. Sun, G. Feng, S. Qin, Y. Liang, and T. P. Yum, "The smart handoff policy for millimeter wave heterogeneous cellular networks," *IEEE Transactions on Mobile Computing*, vol. 17, no. 6, pp. 1456–1468, June 2018.

[22] H. Elshaer, M. N. Kulkarni, F. Boccardi, J. G. Andrews, and M. Dohler, "Downlink and uplink cell association with traditional macrocells and millimeter wave small cells," *IEEE Transactions on Wireless Communications*, vol. 15, no. 9, pp. 6244–6258, Sep. 2016.

[23] G. Ghatak, A. De Domenico, and M. Coupechoux, "Coverage analysis and load balancing in hetnets with millimeter wave multi-rat small cells," *IEEE Transactions on Wireless Communications*, vol. 17, no. 5, pp. 3154–3169, May 2018.

[24] ——, "Modeling and analysis of hetnets with mm-wave multi-rat small cells deployed along roads," in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, Dec 2017.

[25] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular

networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 6, pp. 2706–2716, June 2013.
[26] D. Liu, L. Wang, Y. Chen, T. Zhang, K. K. Chai, and M. Elkashlan, "Distributed energy efficient fair user association in massive mimo enabled hetnets," *IEEE Communications Letters*, vol. 19, no. 10, pp. 1770–1773, Oct 2015.
[27] D. Gesbert, S. G. Kiani, A. Gjendemsjo, and G. E. Oien, "Adaptation, coordination, and distributed resource allocation in interference-limited wireless networks," *Proceedings of the IEEE*, vol. 95, no. 12, pp. 2393–2409, Dec 2007.
[28] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5g cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, 2013.
[29] 3GPP TS 36.211 V8.3.0, " Evolved Universal Terrestrial Radio Access;Physical Channels and Modulation," June 2008.
[30] S. G. Kim, K. Cho, D. Yoon, Y. Ko, and J. K. Kwon, "Performance analysis of downlink inter cell interference coordination in the lte-advanced system," in *2009 Fourth International Conference on Digital Telecommunications*, July 2009, pp. 30–33.
[31] R. W. Rosenthal, "A class of games possessing pure-strategy nash equilibria," *International Journal of Game Theory*, vol. 2, no. 65, 1973.
[32] D. Monderer and L. S. Shapley, "Potential games," *Games and Economic Behavior*, vol. 14, no. 1, pp. 124 – 143, 1996.
[33] M. Grant and S. Boyd. Cvx: Matlab software for disciplined convex programming. [Online]. Available: http://cvxr.com/cvx
[34] J. B. Rosen, "Existence and uniqueness of equilibrium points for concave n-person games," *Econometrica*, vol. 33, pp. 520–534, 1965.
[35] D. P. Palomar, "Convex primal decomposition for multicarrier linear mimo transceivers," *IEEE Transactions on Signal Processing*, vol. 53, no. 12, pp. 4661–4674, Dec 2005.
[36] Mung Chiang, "Balancing transport and physical layers in wireless multihop networks: jointly optimal congestion control and power control," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 1, pp. 104–116, Jan 2005.
[37] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi, "A survey on 3gpp heterogeneous networks," *IEEE Wireless Communications*, vol. 18, no. 3, pp. 10–21, June 2011.
[38] L. Hentila, P. Kyösti, M. Käske, M. Narandzic, and M. Alatossava. Matlab implementation of the winner phase ii channel model ver1.1.
[39] Claussen, "Efficient modelling of channel maps with correlated shadow fading in mobile radio systems," in *2005 IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications*, vol. 1, Sep. 2005, pp. 512–516.
[40] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1164–1179, June 2014.
[41] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive mimo for next generation wireless systems," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 186–195, February 2014.

Kinda Khawam got her engineering degree from Ecole Supérieure des Ingénieurs de Beyrouth (ESIB) in 2002, the Master's degree in computer networks from Telecom ParisTech (ENST), Paris, France, in 2003, and the Ph.D. from the same school in 2006. She was a post doctoral fellow researcher in France Telecom, Issy-Les-Moulineau, France in 2007. Actually, she is an associate professor and researcher at the University of Versailles in France. Her research interests include radio resource management, modeling and performance evaluation of mobile networks. She has co-authored more than 50 papers published in international journals and conference proceedings.

Samer Lahoud is an Associate Professor at the Saint Joseph University of Beirut where he lectures computer networking courses at the Faculty of engineering (ESIB). His research activities focus on routing and resource allocation algorithms for wired and wireless communication networks. He has co-authored more than 80 papers published in international journals and conference proceedings. Mr. Lahoud received the Ph.D. degree in communication networks from Telecom Bretagne, Rennes, in 2006. After his Ph.D. degree, he spent one year at Alcatel-Lucent Bell Labs Europe. From 2007 to 2016, he was with the University of Rennes 1 and with IRISA Rennes as an Associate Professor.

Melhem El Helou (S'08–M'15–SM'19) received the engineer's degree and master's degree in telecommunications and networking engineering from the Ecole Supérieure d'Ingénieurs de Beyrouth (ESIB), Faculty of Engineering at the Saint Joseph University of Beirut, Beirut, Lebanon, in 2009 and 2010, respectively and the Ph.D. degree in computer and telecommunications engineering from IRISA Research Institute, University of Rennes 1, Rennes, France and Saint Joseph University of Beirut, in 2014. He joined ESIB in September 2013 where he is currently an Assistant Professor (*fr:* Maître de conférences). His research interests include wireless networks, radio and energy resource management, Internet of Things, and quality of service.

Steven Martin received his Ph.D. degree from INRIA, France, in 2004. Since 2005, Steven MARTIN is working at Paris-Sud University. Leading the research group "Networking and Optimization" at LRI (the Laboratory for Computer Science at Paris-Sud University, joint with CNRS), his research interests include quality of service, wireless networks, network coding, ad hoc networks and real-time scheduling. He is involved in many research projects and had the lead of the activity "Personal safety in digital cities of the future" in EIT ICT Labs (the European Institute of Innovation and Technology) from 2010 to 2014. He is the author of a large number of papers published in leading conference proceedings and journals. He has served as TPC member for many international conferences in networking.

Gang Feng (M'01–SM'06) received his B. Eng. and M. Eng degrees in Electronic Engineering from the University of Electronic Science and Technology of China (UESTC), in 1986 and 1989, respectively, and the Ph.D. degrees in Information Engineering from The Chinese University of Hong Kong in 1998. He joined the School of Electric and Electronic Engineering, Nanyang Technological University in December 2000 as an assistant professor and was promoted as an associate professor in October 2005. At present, he is a professor with the National Laboratory of Communications, University of Electronic Science and Technology of China. His research interests include resource management in wireless networks, next generation cellular networks, etc. Dr. Feng is a senior member of IEEE.