IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 24, 2018, accepted October 17, 2018, date of publication November 9, 2018,
date of current version November 30, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2877695

Mobility-Based Cell and Resource Allocation for Heterogeneous Ultra-Dense Cellular Networks

MASOUD FAROKHI[1], ALIREZA ZOLGHADRASLI[1], AND NADER MOKARI YAMCHI[2]
[1]Faculty of Electrical and Computer Engineering, University of Shiraz, Shiraz 71348-51154, Iran
[2]Department of Electrical and Computer Engineering, Tarbiat Modares University, Tehran 14115-111, Iran

Corresponding author: Alireza Zolghadrasli (zolghadr@shirazu.ac.ir)

This work was supported by Mobile Communication Company of Iran.

ABSTRACT Heterogeneous ultra-dense network (HUDN) is a promising network structure, which increases network efficiency in 4G and 5G networks. However, it faces the new challenges regarding interference and mobility management. To overcome these challenges, joint of resource allocation (RA) and mobility management is necessary, which, to the best of our knowledge, has not been sufficiently investigated. This paper represents a solution to this issue. First, analytical investigation and numerical analysis are carried out to model and justify the behavior of expectation of handover (HO) success rate ($\mathrm{E}_R\{HSR\}$) versus coverage probability. This paper provides more insight and tools for mobility-based RA researches and design of the network as well. Then, a new approach of hybrid cell-resource allocation is introduced. It is noteworthy that this is a practical structure that is adaptable to dynamic network changes in parameters, such as traffic distribution, mobility pattern, network topology, and different tiers' acceptable signal-to-interference-plus-noise ratio. The advantage of this new proposed approach is demonstrated by a numerical analysis. The results are compared with traditional approaches with and without HO priority consideration called hybrid-partial CRA (HP-CRA) and traditional CRA (T-CRA), respectively. The results show a considerable improvement of $\mathrm{E}_R\{HSR\}$ about 20% and 80% compared with HP-CRA and T-CRA, respectively, under the loaded situations, while the network sum rate is kept near the optimal solution.

INDEX TERMS Hybrid cell allocation and resource allocation, mobility management, stochastic geometry, heterogeneous ultra-dense network, handover success rate, network sum rate.

I. INTRODUCTION

The telecom industry, starting with the mission of connecting people, has been faced with new requests such as internet of things (IoT), Ultra high quality video, etc. [1]. The fact is that it is not only about exponential data traffic and numerous connections growth but also request to high signal to interference signal to interference plus noise ratio (SINR) services such as full HD videos on-demand and online remote surgery. Therefore, proper wireless networks with efficient cell resource allocation (CRA) algorithms should be designed to cope with the related challenges. Heterogeneous ultra-dense network (HUDN) is a promising network structure introduced in 3GPP Rel-10 and recommended for 5G [2], [3]. This structure usually consists of two or three tiers of different types of base stations (BSs), including small cells and macro cells. Small cells are low power nodes which densify the network in randomly-located hotspots. HUDN structure imposes more interference and lower mobility performance compared to homogeneous networks. Considering these challenges for HUDNs, it is necessary to propose practical and efficient mobility-based CRA solutions [4].

A. RELATED WORKS

1) RESOURCE ALLOCATION

Resource allocation (RA) is one functionality of MAC layer in wireless networks such as LTE which has a key role for efficient resource management [5]. This functionality allocates radio resources including subcarriers and power among users such that a utility function, for example network sum rate (NSR) is maximized. NSR is the summation of all users' throughput also called system sum rate [6].

Two types of solutions are generally considered in the literature: instantaneous resource allocation (IRA) [6]–[8] and ergodic resource allocation (ERA) [9]–[11]. In IRA, the instantaneous value of the utility function is optimized by taking instantaneous constraints into account. In accordance with changes in channel state information (CSI), the problem should be solved in a dynamic manner, which introduces high computational complexity and excessive signaling load in network. In ERA, both the utility function and the related constraints are considered in an average sense. ERA solutions are based on the long-term conditions of a wireless system using channel distribution information (CDI). With similar utility and constraint functions, the complexity of ERA is less than that of IRA, but the accuracy of IRA is superior [9]. In [10], a simplified hybrid RA solution for spectrum allocation as offline phase and power allocation as online phase is presented. However, cell allocation and mobility management are not considered.

2) MOBILITY MODELING

In order to analyze and simulate the mobility performance, we should consider a proper mobility model to describe the movement pattern of mobile users in cellular networks. Several models have been presented which can be divided into two groups: dependent models [12] and random mobility models [13]. In dependent models, the mobility parameters depend on geography, location, direction, etc. This type of analysis is recommended for benchmarking and analysis in specific situations. However, it is not proper for general cases and analytical investigations. In random models, characteristics such as direction and velocity follow random process. In this regard, Random waypoint (RWP) model is a very well-known and tractable model [13], [14]. Moreover, this is a generic model which can be easily used for mathematical and numerical analysis.

3) MOBILITY-BASED RA

RA and mobility management play key roles for improving network performance. It is worth noting that since they are dependent, they should be considered together in a unified framework. This issue is addressed by [15] and [16]. It is demonstrated through simulation that mobility and interference are interwoven. In [17], a RA solution with the target of minimizing energy is proposed. This method categorizes mobile users into different groups and analyzes SINR for each group. However, no mobility criteria such as handover success rate (HSR) is investigated. Generally, HSR describes the ratio of the number of successfully performed handover procedures to the number of attempted handover procedures [18]. Reference [19] presents a scheduling method focusing on multimedia services using separate handover (HO) control and RA modules. The efficient throughput called good put is simulated. However, mobility is not investigated. Reference [14] introduces mobility predictive method to improve network resource utilization. This study analyzes the performance of network from capacity point of view and dose not tackle the performance indicators of mobility such as HSR.

5G mobility-aware user associations is presented in [20], in order to overcome the limitations of the conventional received power (RSS)-based association strategies. It proposes an optimization problem which is helpful in dynamic situations caused by users' mobility and BSs' loads. However, the solution is limited to mmWave systems without considering the interference and allocating radio resources such as transmit power and subcarrier.

As mentioned above, there are few solutions presenting joint cell allocation and RA which take into account mobility performance, coverage, load of the network, and flexibility to dynamic changes of network. One of the main 4G HUDNs' challenges is the destructive impact of mobility on performance which is more serious in 5G [4]. Accordingly, as one of the main prerequisite to reach the targets of 5G, we should optimize the process and algorithms of RA and mobility management jointly.

This paper presents H-CRA, a new kind of joint CRA algorithm, which satisfies both coverage and mobility performance requirements. H-CRA improves the efficiency of CRA to serve both stationary and mobile users in the context of dynamic HUDN with high SINR service requirements. Joint process of CRA and mobility does not mean removing HO management procedure. However, it considers both coverage and HSR requirements to allocate resources in a more efficient way and removes some unnecessary round trips between CRA and HO algorithms.

B. MAIN CONTRIBUTIONS

i- Mobility in HUDNs is modeled by exploiting the RWP method and stochastic geometry. Through analytical and numerical investigation, we obtain an interesting result showing that the geographical expectation of HSR ($\mathrm{E}_R\{HSR\}$) in HUDNs is equivalent to the coverage probability.

ii- To the best of our knowledge, no research presents CRA problems which takes into account the mobility behavior of users in terms of HSR. Accordingly, the proposed RA covers both stationary and mobile subscribers and the HSR factor can be improved considerably.

iii- The optimization problem is proposed as a combination of cell allocation and resource allocation to enhance coverage and mobility situations. In this regard, we model coverage probability by using stochastic geometry.

iv- A new approach of hybrid CRA consisting of offline and online phases is introduced. This approach has a practical and generic structure for CRA which is recommended in complex and dynamic situations of HUDNs. In the offline phase, an ERA problem is solved which satisfies the constraints in average sense such as $\mathrm{E}_R\{HSR\}$ and coverage probability. Although the proposed solution is time-consuming, it can only be solved for long-term conditions of users and network. This phase determines initial transmit power values and bandwidth for mobile users which should be used in the online phase. Then, an IRA problem as online phase is solved which instantaneously fulfills the constraints at each time slot.

These derivations and techniques facilitate analysis and resource management more efficiently due to the following advantages:

- The results provide applicable insights about service design at each tier.
- This approach reduces the load of BSs caused by ping pong cell change or HO due to the load of BSs or low quality of connections.
- It also provides a flexible tool to track the dynamic changes of the network parameters including number and locations of BSs, threshold of SINR, the models of users' mobility, and traffic model.
- Eventually, it improves the most important performance indicator of mobility management i.e., HSR.

In current wireless systems, the main part of HO algorithms belongs to the radio resource management layer (L3). Moreover, resource allocation is performed in the MAC layer separately from HO but interactively with L3. The focus of this paper is on the MAC layer mobility-based resource allocation to enhance the overall mobility management performance.

In order to investigate H-CRA in more details, hybrid-partial CRA (HP-CRA) and traditional CRA (T-CRA) are presented. HP-CRA is a special solution of H-CRA with near-optimal results of NSR. However, H-CRA outperforms HP-CRA.

The organization of the paper is as follows: Section II introduces the models deployed in this paper. Section III derives the mathematical relations of coverage probability and $\mathrm{E}_R\{HSR\}$. The analytical result of $\mathrm{E}_R\{HSR\}$ is justified through numerical analysis. Section IV discusses the proposed H-CRA, HP-CRA and T-CRA to compare their performance. Section V presents numerical analysis and performance evaluation. Section VI is dedicated to the conclusion. The proof of Lemma and the list of abbreviations are presented in the appendix section.

II. SYSTEM MODEL

A. NETWORK MODEL

This study is devoted to the downlink of an Orthogonal Frequency Multiple Access (OFDMA)-based HUDN. It is assumed that the HUDN network consists of I tiers belonging to set $\boldsymbol{I} = \{1, \cdots, I\}$ where each tier $i \in \boldsymbol{I}$ specifies one type of BS, such as macro, micro, pico or small cell. The main specifications such as transmit power, capacity, and density of the BSs may differ in each tier. Furthermore, it is assumed that these cells belong to open access category. The considered HUDN structure is modeled based on stochastic geometry. Stochastic geometry is a well promising technique which has been used vastly in the recent years [21]–[25]. It is a general tool which is much more tractable than the grid model for analysis and simulation of the cellular networks especially HUDNs. Users in cellular networks also small cells in heterogeneous network are located randomly which are not possible to be modeled by the grid or other deterministic models. Therefore, we have to adopt a random model such as stochastic geometry to study behaviors of HUDNs including coverage probability and mobility. The reason is that the stochastic geometry allows us to study the average behavior over many spatial realizations of a network whose nodes (BSs or users) are placed according to some probability distributions.

One of the important objects of stochastic geometry is point process model which is the random collection of points in the space and can be chosen from different types. Poisson Point Process (PPP) is the most usual and typical point process which is proper to model the locations of BSs and users. PPP provides tractable tool for mathematical analysis with acceptable accuracy to model practical cases [26].

In this regard, the BSs are spatially distributed in each tier using PPP ϕ_i of density λ_i. The average transmit power of each BS in tier i is indicated by $P_{Tier(i)}$ and the SINR threshold of each BS in tier i is denoted by β_i. Similarly, the distribution of users is modeled by independent PPP ϕ_m with density λ_m.

Considering Rayleigh fading channel, the channel power gain is denoted by $\{h^m_{f,n}\colon \exp(1)\}$ related to the connection between the m^{th} user and BS or cell $f \in \{1, 2, \cdots, F\}$ on subcarrier n. $\exp(1)$ indicates the exponential function with the expected value one [27] and $x_{f,m}$ stands for the distance between user m and BS f. The power loss is $l(x) = \|x\|^{-\alpha}$, where $\alpha \geq 2$ is the path loss exponent and $\|\cdot\|$ is the norm function. The total bandwidth of the network is divided into N orthogonal subcarriers, denoted by $n \in \mathrm{N} = \{1, 2, \cdots, N\}$ and each of them undergoes block fading.

We consider $\boldsymbol{P}$ as the related transmission power matrix where its elements are denoted by $p^m_{f,n}$. $P_{Tier(i)}$, the average transmit power of base stations in tier i, belongs to $\boldsymbol{P}_{Tier}$ with the dimension of $1 \times I$. $\rho^m_{f,n} \in \{0, 1\}$ is a binary variable showing the result of subcarrier allocation for the probable user $m \in \{1, 2, \cdots, M\}$ in cell f. $\rho^m_{f,n}$ is the element of $\boldsymbol{\rho}$ where its dimension is $F \times M \times N$. We define matrix $\boldsymbol{B}$ with the dimension of $M \times F$ with element b^m_f representing if user m is assigned to the cell f. The users are classified into two sets of mobile and stationary users denoted by M_m and M_s respectively. The notations are summarized in Table. 1. The SINR of the m^{th} user connecting to cell f on subcarrier $n \in \mathrm{N}$ is indicated by $\gamma^m_{f,n}$. $\gamma^m_{f,n}$ is calculated as follows:

$$\gamma^m_{f,n} = \frac{\rho^m_{f,n} p^m_{f,n} h^m_{f,n} \left\|x_{f,m}\right\|^{-\alpha}}{\sigma^2 + \sum_{j=1, j\neq m}^{M} \sum_{k=1, k\neq f}^{F} \rho^j_{k,n} p^j_{k,n} h^m_{k,n} \left\|x_{k,m}\right\|^{-\alpha}}, \tag{1}$$

where σ^2 is the constant additive noise power. Without loss of generality, it is assumed that the CSI is perfectly available and backhaul links have sufficient bandwidth. We will comment more on this assumption in section IV, discussion 4. In this paper, each macro BS, as first tier, controls the overlaid BSs lied in other tiers. Scheduling is performed in the macro BS using its own measurement and those from the overlaid and neighbor BSs. The macro BSs exchange the measurements and resource allocation information with each other to overcome interference across different tiers. Two samples of the three-tier HUDN are shown in Fig. 1 as a mix of macro, micro, and pico cells. These BSs are located based on aforementioned PPP model. The area is tessellated by the

TABLE 1. Notations description.

Notation	Description
$i \in \boldsymbol{I} = \{1,\cdots,I\}$	Index of tiers
$\phi_i \in \Phi$	BSs PPP in tier i
$\lambda_i \in \boldsymbol{\lambda}$	BSs PPP density in tier i
$P_{Tier(i)} \in \boldsymbol{P_{Tier}}$	average BSs transmit power in tier i
$\beta_i \in \boldsymbol{\beta}$	SINR threshold of each BS in tier i
ϕ_m	Users PPP
λ_m	Users PPP density
$h_{f,n}^m \in \mathrm{H}$	Channel power gain
$n \in \mathrm{N} = \{1,2,\cdots,N\}$	Subcarrier set
$\rho_{f.n}^m \in \boldsymbol{\rho}$	Subcarrier allocation indicator
$m \in \{1,2,\cdots,M\}$	Users set
$b_f^m \in \boldsymbol{B}$	Cell allocation indicator
M_m	Mobile users set
M_s	Stationary users set
$f \in \{1,2,\cdots,F\}$	Cells set
$p_{f,n}^m \in \boldsymbol{P}$	BSs' power
$\gamma_{f,n}^m$	SINR of user m 's connection to cell f on subcarrier n
$\mathcal{P}_c$	Coverage probability
$x_{f,m}$	distance between user m and BS f
$r_{f,n}^m$	The data rate of user m on subcarrier n served by BS f
$\mathcal{P}_{min}$	Coverage probability threshold
τ	HSR probability threshold
ξ	$\max\{\mathcal{P}_{min},\tau\}$
$\boldsymbol{\theta}^*$	Cells' bandwidth reserved for incoming HO users

average coverage regions in the absence of fading. They are obtained by taking average over a period of time based on maximum SINR connectivity model which is defined in the next section.

B. MOBILITY MODEL

As mentioned in the introduction, mobility models are generally divided into two types, namely random models and dependent models. In random models, the parameters such as velocity, direction, etc. are based on random processes. While the parameters in dependent models are function of geography, location, direction, and clustering, etc. Since our main aim is to obtain an analytical relations and model for general mobility management, it is not proper to utilize the dependent model. Therefore, we adopt the random model without loss of generality. In this regard, Random Way Point (RWP) is the most useful random model which is widely used for both analytical studies and simulations [28], [29].

RWP is defined as a sequence of triples within a convex two-dimensional area $\mathcal{A} \subset \mathbb{R}^2$:

$$(X_1, X_2, V_1), (X_2, X_3, V_2), (X_3, X_4, V_3), \cdots \tag{2}$$

where $\{X_1, X_2, X_3\}$ are random points selected from both uniform and non-uniform distributions over $\mathcal{A}$. In order to verify the random model's results, non-uniform distribution of RWP which is somehow close to the dependent models is investigated in the paper. Every two points in the triples constitute a leg in which the user moves along that leg with average velocity V_i. Fig. 1 shows two samples of mobility pattern with a limited number of legs.

III. COVERAGE PROBABILITY, HANDOVER SUCCESS RATE

A. DEFINITION: GENERAL COVERAGE PROBABILITY

A user mobile is defined to be in coverage if there exists at least one connection to each cell f across the tiers where the SINR is greater than or equal to threshold β_i. Considering CRA algorithm which assigns different subcarriers to each user, the coverage condition is satisfied if at least one subcarrier n has a SINR greater than or equal to β_i. Therefore, the coverage probability of a randomly located user m is formulated as:

$$\mathcal{P}_c(\{\boldsymbol{\lambda}\}\{\boldsymbol{\beta}\}\{\boldsymbol{P_{Tier}}\}) = \mathbb{P}\left\{\bigcup_{i\in I}\bigcup_{x_f\in\Phi_i}\bar{\gamma}_f^m \geq \beta_i\right\}, \tag{3}$$

where $\boldsymbol{P}_{Tier}$ is the average transmit power of base stations in different tiers and $\bar{\gamma}_f^m$ is the maximum of $\gamma_{f,n}^m$ versus N subcarriers. $\mathbb{P}\{.\}$ is the probability function.

Using the results of [21], the coverage probability is:

$$\mathcal{P}_c = \sum_{i=1}^{I}\lambda_i\int_{R^2}\exp[-\left(\frac{\beta_i}{P_{Tier(i)}}\right)^{2/\alpha}C(\alpha)\left\|x_{f,m}\right\|^2 \times\sum_{j=1}^{I}\lambda_j\left(P_{Tier(j)}\right)^{2/\alpha}]\exp\left(-\frac{\beta_i\sigma^2}{P_{Tier(i)}}\left\|x_{f,m}\right\|^{\alpha}\right) \times dx_{f,m} \tag{4}$$

where $C(\alpha) = 2\pi^2\csc\left(\frac{2\pi}{\alpha}\right)\alpha^{-1}$. ■

Nowadays, mobile services' throughput requirement needs the SINR to be much more than one. Moreover, HUDNs are usually interference limited due to the limited shared spectrum among the tiers. Hence, it is very helpful to utilize the simplification of (4). Based on [21, Corollary 1] and by assuming the interference limited networks with negligible noise and $\beta_i \geq 1$ ($0dB$), (4) is simplified to:

$$\mathcal{P}_c = \frac{\pi}{C(\alpha)}\frac{\sum_{i=1}^{I}\lambda_i P_{Tier(i)}^{2/\alpha}\beta_i^{-2/\alpha}}{\sum_{i=1}^{I}\lambda_i P_{Tier(i)}^{2/\alpha}} \tag{5}$$

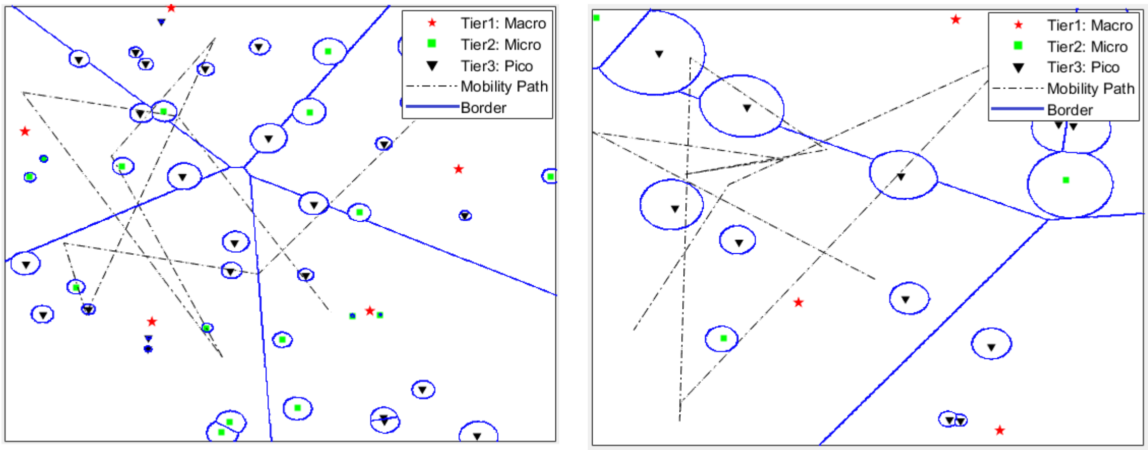


FIGURE 1. Two close-up views of 3-tier HUDN with different settings of BS density and path loss exponent. Left: $\lambda = [7e^{-6}, 14e^{-6}, 28e^{-6}]$; $\alpha = 3$, Right: $\lambda = [1e^{-6}, 5e^{-6}, 10e^{-6}]$; $\alpha = 4$.

Discussion 1: In Fig. 2, the coverage probability behaviors are drawn versus the first and second tiers' average transmit powers $\{P_1, P_2\}$ considering different configurations. The graphs show that the P_i which maximizes $\mathcal{P}_c$ is varying depending on the value of β_i. Another important point is that the behavior of the coverage probability is monotonic which allows us to deploy different algorithms for feasible optimization problems.

The following Proposition introduces an interesting fact when all tiers have the same SINR threshold:

Proposition 1: Coverage probability $\mathcal{P}_c$ in an interference limited network is independent of tiers' average power $\boldsymbol{P}_{Tier}$ when β_is are the same for all $i \in \{1, \cdots, I\}$ and greater or equal than one.

Proof: Taking into account (5), it is evident that when for all tiers $\beta_i = \beta$ then $\mathcal{P}_c^* = \frac{\pi}{C(\alpha)}\beta^{2/\alpha}$. ■

B. DEFINITION: HSR

In this paper, HSR is modeled mathematically as the probability of successful handover in a convex area $\mathcal{A}$. In fact, HO is successful when the moving user m has at least one connection to any tier with SINR which is greater than or equal to the threshold (β_i). Alternatively, handover failure is happened when there is no connection to tier $i \in I$ for moving user m such that its connection's SINR is greater than or equal to the required thresholds (β_i).

According to the random behavior of the users and the networks, we adopt the geographical expectation of HSR, $\mathrm{E}_R\{HSR\}$, or alternatively $\mathrm{E}_R\{HFR\}$ in our analysis.

C. LEMMA 1

Assuming a typical mobile user in a free capacity HUDN, the geographical expectation of handover success rate is equal to the coverage probability regardless of the mobile users' location PDF

$$\mathcal{P}_c = \mathrm{E}_R\{HSR\}. \tag{6}$$

Proof: See Appendix.

Discussion 2: In order to justify the mobility modeling and the results of Lemma 1, comprehensive simulation is performed which verifies the theoretical results. Fig. 3 presents $\mathcal{P}_c$ and $\mathrm{E}_R\{HSR\}$ versus β_1 in order to compare the proposed analytical and simulation results. Both uniform and non-uniform distributions of mobile users are investigated. In this figure, density of BSs across the tiers are $\boldsymbol{\lambda} = \left[1e^{-4}, 5e^{-4}, 15e^{-4}\right]$. In the first scenario, $\beta_i = 1$ for $i = 2, 3$ while in the second scenario, $\beta_i = 0.5$ for $i = 2, 3$. As shown in this figure, the analytical and simulation results are very closed which verify Lemma 1.

Discussion 3: Every relation of $\mathcal{P}_c$ which is based on definition (3) can be used in Lemma 1. Equation (4) is general and closed form relation of $\mathcal{P}_c$ which is valid for a wide range of SINR from high values down to -4 dB even under weaker assumptions [21]. Furthermore, it should be noted that in many cases, HUDNs are interference limited and noise is negligible. By numerical investigations, we reach to this fact that the behavior of (4) and (5) are very similar as indicated in Fig. 3. Another point is that regarding the high penetration of smart phones and demand for high performance services, the condition $\beta_i \geq 1$ is a common threshold of coverage for the new applications. Thus, (5) is adopted in the next sections as $\mathcal{P}_c$.

IV. HYBRID JOINT CRA PROBLEM

In this section, we propose a general network-level problem consisting of three sub-problems: Cell allocation which assigns users to the best cells and RA including both transmit power and subcarrier allocation sub-problems.

The main goal of this optimization problem is to maximize the network sum rate (NSR) taking into account the coverage conditions for stationary users and successful HO conditions for mobile users in average sense. Accordingly, two constraints are included in the optimization problem. One constraint controls CRA solution in such a way that the coverage

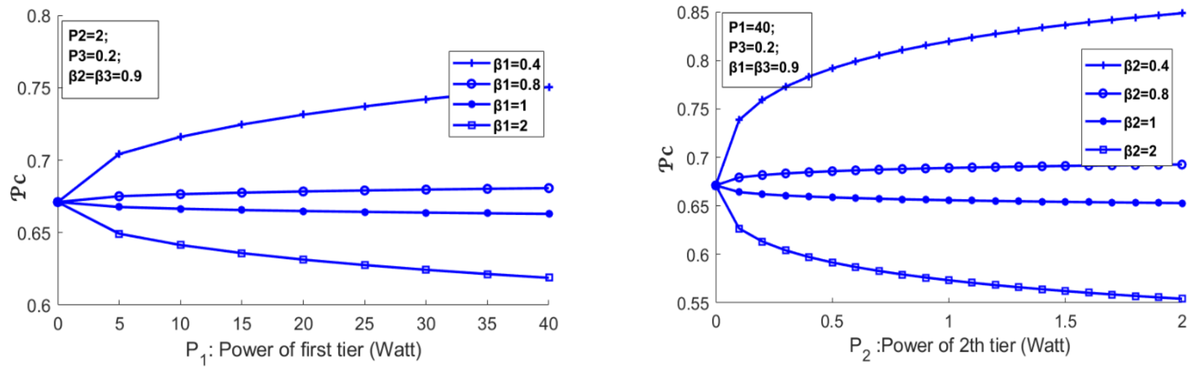


FIGURE 2. Coverage probability versus P_1 and P_2 when $\lambda = [1e^{-6}, 1e^{-5}, 1.5e^{-5}]$ based on two scenarios with different β_i.

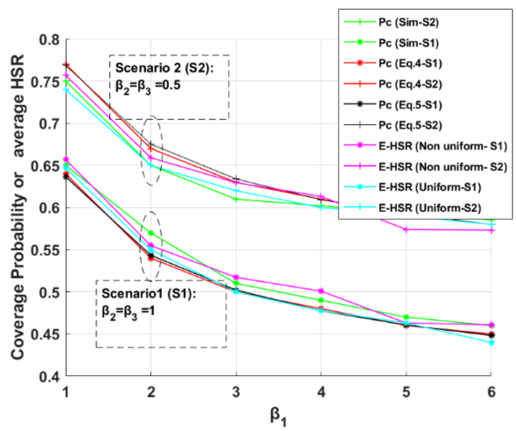


FIGURE 3. Comparison of $\mathcal{P}_c$ and normalized $E_R\{HSR\}$ for the considered methods: analytical relations (4), (5), simulation of $\mathcal{P}_c$, and simulation of $E_R\{HSR\}$ with both uniform and non-uniform PDF of mobile users. Two settings of SINR sensitivity are considered which are denoted by S1 and S2.

probability is greater than or equal to the threshold ($\mathcal{P}_{min}$). Another constraint is the average HSR should be greater than or equal to the threshold (τ). The thresholds τ and $\mathcal{P}_{min}$ are the normalized variables in the range of (0, 1]. Moreover, it is assumed that the network is heterogeneous cellular network which is modeled based on the method denoted in Section II. Note that some illustrations of the coverage probability and HSR are provided in Fig. 2 and Fig. 3.

Moreover, maximum transmit power of each BS and subcarrier allocation constraints are considered. Hence, the proposed optimization problem is written as follows:

$$\text{Target:} \max_{\boldsymbol{P},\boldsymbol{B},\boldsymbol{\rho}} \{NSR\} = \max_{\boldsymbol{P},\boldsymbol{B},\boldsymbol{\rho}} \times \left\{\sum_{f=1}^{F}\sum_{m=1}^{M}\sum_{n=1}^{N} b_f^m \rho_f^m \log_2(1+r_{f,n}^m)\right\}, \tag{7a}$$

$$\text{s.t } E_R\{HSR\} \geq \tau; \quad \forall m \in M_m \tag{7b}$$

$$\mathcal{P}_c \geq \mathcal{P}_{min}; \quad \forall m \in M_s \tag{7c}$$

$$\rho_{f,n}^m \in \{0,1\}; \quad \forall m, f, n \tag{7d}$$

$$\sum_{m=1}^{M} \rho_{f,n}^m \leq 1; \quad \forall n, f \tag{7e}$$

$$\sum_{m=1}^{M}\sum_{n=1}^{N} \rho_{f,n}^m p_{f,n}^m \leq P_f^{max}; \quad \forall f \tag{7f}$$

$$b_f^m \in \{0,1\}; \quad \forall m, f \tag{7g}$$

$$\sum_{f=1}^{F} b_f^m \leq 1; \quad \forall m \tag{7h}$$

where $r_{f,n}^m$ is defined as the data rate of user m on subcarrier n served by BS f:

$$r_{f,n}^m = \log_2\left(1+\gamma_{f,n}^m\right) \tag{8}$$

(7b) represents the constraint of the minimum acceptable mean of HSR for HO users and (7c) is the coverage probability constraint for stationary users. (7e) states that each subcarrier cannot be shared by two or more users in one cell at the same time. The transmit power of the BS should be less than or equal to P_f^{max} formulated in (7f). (7h) limits each user to be served only by one BS.

Since the problem is non-convex and mixed integer-nonlinear, it is too complex to be solved optimally. However, there are some facts which are worthy of note: the variables such as network structure, tiers' SINR sensitivity, β_i, and even overall network mobility pattern are not changing very fast in practice during some time intervals such as 10 ms LTE-frame's duration [5].

Consequently, we exploit this fact to divide the problem into two parts: offline, called E-CRA, and online called I-CRA, in such a way to fulfill the target function and all constraints in (7).

Based on the aforementioned points, the ergodic sense of objective and constraint functions are adopted in the offline phase. This phase is again divided into two parts: First, average transmit power of the tiers, $\boldsymbol{P}_{Tier}^*$ is calculated. Then, the needed bandwidth of each BS to be reserved for mobility is obtained which is denoted by $\boldsymbol{\theta}^*$. $\boldsymbol{P}_{Tier}^*$ and $\boldsymbol{\theta}^*$ are $1 \times I$ and $1 \times F$ matrices, respectively.

Unlike the offline phase, I-CRA is designed in the instantaneous sense considering all cells, subcarrier and power allocation parts. Fig. 4 summarizes the proposed H-CRA approach.

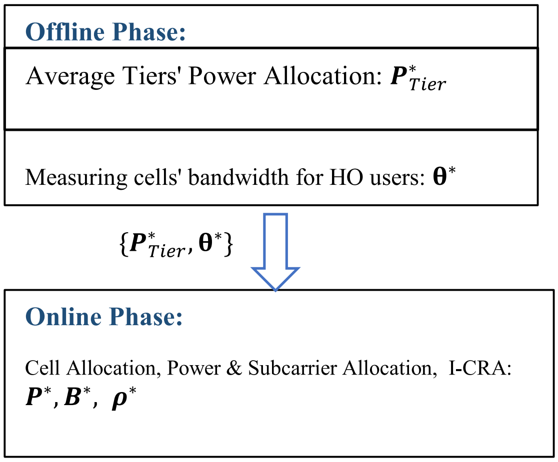


FIGURE 4. Overall structure of H-CRA solution and optimization variables. The star indices stand for the sub-optimal solutions.

A. OFFLINE PHASE (E-CRA)

The offline problem is ergodic and complex. However, handling the offline problem during a more relaxed time period allows us to cope with this complexity. Regardless of complexity, it is also needed to analyze the behavior of the network in a longer time than a short time as time slot. Therefore, we design the offline problem in order to obtain: a) optimum average transmit power of each tier, and b) the bandwidth needed to be reserved for mobile users or HO at each cell.

Propopsition 2: In order to solve (7), considering the ergodic sense of the offline problem and the results of Lemma 1, the constraints (7b), (7c) are equivalent and merged as:

$$\mathcal{P}_c \geq \xi; \tag{9}$$

where $\xi = \max\{\mathcal{P}_{min}, \tau\}$.

1) OFFLINE PHASE, PART 1

Optimum values of $\boldsymbol{P}_{Tier}$ is obtained by using (7) and coverage probability behavior presented in Section 3 and Proposition 2. Then it is possible to solve the problem in the offline mode using expectation of NSR, E$\{NSR\}$, as our target function in an ergodic optimization problem. Interestingly, if we note that almost all HUDNs are interference limited due to the usage of the same limited spectrum in all tiers and $\boldsymbol{\beta} \succcurlyeq \mathbf{1}$ which is the proper threshold in real networks such as LTE then, we can simplify E$\{NSR\}$ to average user throughput $\bar{R}$ if the user is in coverage [21]:

$$\bar{R} = \log\left(1 + \beta_{min}\right) + \frac{\sum_{i=1}^{I} \lambda_i P_{Tier(i)}^{2/\alpha} \beta_i^{-2/\alpha} A\left(\alpha, \beta_i, \beta_{min}\right)}{\sum_{i=1}^{I} \lambda_i P_{Tier(i)}^{2/\alpha}}. \tag{10}$$

where $A\left(\alpha, \beta_i, \beta_{min}\right) = \int_{\beta_{min}}^{\infty} \frac{\max(\beta_i, x)^{-2/\alpha}}{1+x} dx$ and $\beta_{min} = \min\{\boldsymbol{\beta}\}$.

Hence, the target is to maximize $\bar{R}$ subject to constraints (7b) and (7c) where the maximum transmit power of each tier i is denoted by $P_{Tier(i)}^{max}$. The solution of this part enables us to determine the priority of transmit power among all tiers and initial values for online power allocation.

$$\text{Target: } \max_{\boldsymbol{P}_{Tier}} \left\{\bar{R}\right\}, \tag{11a}$$
$$\text{s.t: } \mathcal{P}_c \geq \xi; \tag{11b}$$
$$P_{Tier(i)} \leq P_{Tier(i)}^{max}; \quad \forall i \tag{11c}$$

The Lagrangian function of problem (11) is given by:

$$L_1\left(\boldsymbol{P}_{Tier}, \boldsymbol{\mu}\right) = \bar{R} + \mu_{1,1}\left(\mathcal{P}_c - \xi\right) + \sum_{i=1}^{I} \mu_{2,i}(P_{Tier(i)}^{max} - P_{Tire(i)}). \tag{12}$$

where $\boldsymbol{\mu}$ is the Lagrange multiplier vector. In order to find the optimum value, we solve $\frac{\partial L_1}{\partial P_{Tire(i)}} = 0$ which yields the following polynomial equation. It can be solved numerically versus other tiers' powers by considering fixed values of Lagrange multipliers.

$$\mathrm{K}_1 P_{Tier(i)}^{\frac{2}{\alpha}-1} - \left(\lambda_i P_{Tier(i)}^{\frac{2}{\alpha}} + \mathrm{K}_2\right) = 0, \tag{13}$$

where $\mathrm{K}_1 = \frac{2(A+\mu_{1,1})\lambda_i}{\alpha\mu_{2,i}} \sum_{j\neq i}^{I} \lambda_j P_{Tier(j)}^{\frac{2}{\alpha}}(1 - \beta_j^{-\frac{2}{\alpha}})$ and $K_2 = \left(\sum_{j\neq i}^{I} \lambda_j P_{Tier(j)}^{\frac{2}{\alpha}}\right)^2$.

Considering that HUDNs are almost located in dense urban and hotspots areas, we can assume that path loss $\alpha = 4$ [27]. After some manipulations, the closed-form solution of (13), $\boldsymbol{P}^*_{Tier}$ is obtained as follows:

$$\boldsymbol{P}^*_{Tier} = \frac{\vartheta_i}{3.78\lambda_i^4} - \frac{3.78\left(12\mathrm{K}_1\mathrm{K}_2\lambda_i^5 - \mathrm{K}_2^4\lambda_i^4\right)}{3\lambda_i^4\vartheta_i} + \frac{2\mathrm{K}_2^2}{3\lambda_i^2}, \tag{14a}$$

where ϑ_i is:

$$\left(\sqrt{255.4K_1^3K_2^3\lambda_i^{15} + 283K_1^2K_2^6\lambda_i^{14} + 36.5K_1K_2^9\lambda_i^{13}} + 27K_1^2\lambda_i^8 - 72K_1K_2^3\lambda_i^7 - 2K_2^6\lambda_i^6\right)^{1/3}. \tag{14b}$$

Lagrangian problem (12) is solved by exploiting the dual optimization and sub-gradient algorithms as follows

$$g_1(\boldsymbol{\mu}) = \max_{\boldsymbol{\mu}} L_1\left(\boldsymbol{P}^*_{Tier}, \boldsymbol{\mu}\right) \tag{15}$$

Considering the dual optimization problem:

$$\min_{\boldsymbol{\mu}} \; g_1(\boldsymbol{\mu}), \tag{16a}$$
$$\text{s.t } \boldsymbol{\mu} \geq \mathbf{0}, \tag{16b}$$

the Lagrange multipliers $\boldsymbol{\mu}$ are determined using the sub-gradient method [30] as:

$$\mu_{1,1}^{l+1} = [\mu_{1,1}^{l} - s^l\left(\mathcal{P}_c - \xi\right)]^+; \tag{17a}$$
$$\mu_{2,i}^{l+1} = [\mu_{2,i}^{l} - s^l\left(P_i^m - P_{Tier(i)}\right)]^+; \quad \forall i \tag{17b}$$

where l is the iteration number and $[\cdot]^+$ denotes max (., 0).
where l is the iteration number and $[\cdot]^+$ denotes max (., 0).

s^l is the scalar step size which has a key role in the convergence of (17). It is demonstrated that the sub-gradient approach in the dual decomposition method converges to optimal μ if s^l is chosen to be sufficiently small. As a proper criterion, s^l is selected to be square summable, and it should not be absolute summable [26].

Discussion: Fig. 2 indicates $\mathcal{P}_c$ behavior versus different values of $\boldsymbol{\beta}$. In some cases, $\boldsymbol{P}^*_{Tier}$ happens in higher values of β_is and in some situations vice versa. However, if β_is are equal for all tiers then $\boldsymbol{\lambda}$ and transmit power $\boldsymbol{P}_{Tier}$ does not affect the value of $\mathcal{P}_c$.

It follows from Fig. 2 and the behavior of $\bar{R}$ [21] that (11) is concave. Subsequently, considering Karush_Kuhn_Tucker (KKT) [31] as the first optimality condition, depicted in Table. 2, it is feasible to attain the optimum solution, $\boldsymbol{P}^*_{Tier}$.

TABLE 2. KKT conditions of problem (11) where P^*_{Tier} is the solution.

1.	The related constraints to power (11b) and (11c) must be fulfilled for $\boldsymbol{P}^*_{Tier}$
2.	$\boldsymbol{\mu} \geq 0$
3.	$\mu_{1,1}\left(\mathcal{P}_c - \xi\right) \geq 0$
4.	$\mu_{1,i}\left(P_i^{max} - P_{Tier(i)}\right) \geq 0;$
5.	$\nabla \mathrm{L}_1(\boldsymbol{P}, \boldsymbol{\mu}) = 0$

Insight: A supplementary application is to deploy the optimization problem (11) and its solution in network design process. For instance, if the solution results to $\boldsymbol{P}^*_{Tier(3)} = 0$, then we find that the third tier is not useful versus $\mathcal{P}_c$ and may be applicable only for capacity. If so, it is reasonable to revise the service strategy in order to change the SINR sensitivity of the 3$^{\text{th}}$ tier. Hence, $\mathcal{P}_c$ and user's experience can be improved.

2) OFFLINE PHASE, PART2

In the call admission and CRA process, two sets of stationary and mobile users request resources. In the proposed approach, it is required that mobile users' online CRA to be carried out in parallel with stationary users' CRA, assuming pre-determined capacity at each cell. In order to calculate this capacity for mobile users, the number of requested HOs for each cell is measured during the offline time. Therefore; the mean number of needed subcarriers for HO admission called $\boldsymbol{\theta}^*$ is produced and transferred to online problem.

B. ONLINE PHASE

The online phase is a complementary part to the offline phase and enhances the accuracy of the overall solution. This phase is devoted to find the solutions to the cell, subcarrier, and power allocation problems in an instantaneous sense based on the results of the offline phase.

This problem is a combination of three sub-problems: power, cell, and subcarrier allocation sub-problems. Based on the alternative search method (ASM) [32] and extending the result of [33], these stages are iteratively solved to converge to the best sub-optimal solution.

$$\begin{aligned}
&\{\boldsymbol{P}[0] \rightarrow \boldsymbol{B}[0] \rightarrow \boldsymbol{\rho}[0]\} \\
&\quad \Longrightarrow \cdots \Longrightarrow \{\boldsymbol{P}[t_{opt-1}] \rightarrow \boldsymbol{B}[t_{opt-1}] \rightarrow \boldsymbol{\rho}[t_{opt-1}]\} \\
&\quad \Longrightarrow \{\boldsymbol{P}[t_{opt}] \rightarrow \boldsymbol{B}[t_{opt}] \rightarrow \boldsymbol{\rho}[t_{opt}]\}. \qquad (18)
\end{aligned}$$

where $\{\boldsymbol{P}[0], \boldsymbol{B}[0], \boldsymbol{\rho}[0]\}$ is the initial feasible set of CRA where $\boldsymbol{P}[0]$ follows $\boldsymbol{P}^*_{Tier}$ i.e., $p^m_{n,f} = P^*_{Tier(f)}/N$. Then, at the beginning of each iteration t, $\boldsymbol{P}[t]$ is computed versus $\boldsymbol{B}[t-1]$ and $\boldsymbol{\rho}[t-1]$ obtained in the last iteration. These iterations are repeated until convergence is achieved. The overall process of online solution is summarized in Table. 3.

TABLE 3. Online cell & resource allocation algorithms.

Cell-Subcarrier and Power Allocation
Initialize t :=0, $\boldsymbol{B}[\boldsymbol{0}], \boldsymbol{\rho}[\boldsymbol{0}], \boldsymbol{P}^*_{Tier}$
repeat To solve online (18)
For fixed $\boldsymbol{P}[t-1]$, find $\boldsymbol{\rho}[t]$ and $\boldsymbol{B}[t]$ using (24) and (25).
For fixed $\boldsymbol{\rho}[t]$ and $\boldsymbol{B}[t]$, calculate $\boldsymbol{P}[t+1]$ using $\boldsymbol{P}[t]$ and solving $\frac{\partial L_2}{\partial P} = 0$, (21)
For fixed $\boldsymbol{P}[t+1]$, find $\boldsymbol{v}$ using the sub-gradient method proposed in (22)
Set t := t + 1
until convergence of (22)
Output: $\boldsymbol{\rho}, \boldsymbol{B}$ *and* $\boldsymbol{P}$

1) ONLINE POWER ALLOCATION

Similar to the main problem (7), NSR is chosen to be the objective function in the instantaneous sense meaning that it is carried out in each time slot. The online power allocation sub-problem is formulated as follows:

Online Target:

$$\max_{\boldsymbol{P}} \sum\nolimits_{m=1}^{M} \sum\nolimits_{n=1}^{N} \sum\nolimits_{f=1}^{F} b_f^m \rho_{f,n}^m r_{f,n}^m, \qquad (19a)$$

$$\text{s.t:} \sum\nolimits_{f=1}^{F} \sum\nolimits_{m=1}^{M} \rho_{f,n}^m \leq 1; \quad \forall n \qquad (19b)$$

$$\sum\nolimits_{m=1}^{M} \sum\nolimits_{n=1}^{N} \rho_{f,n}^m p_{f,n}^m \leq P_f^{max}; \quad \forall f \qquad (19c)$$

$$\rho_{f,n}^m \in \boldsymbol{\theta}^*; \quad \forall m \in M_m, \ \forall f \qquad (19d)$$

$$b_f^m \in \{0, 1\}; \quad \forall m, f \qquad (19e)$$

$$\sum\nolimits_{f=1}^{F} b_f^m \leq 1; \quad \forall m \qquad (19f)$$

Due to the non-convexity of objective function in (19), we can exploit the successive convex approximation with low complexity (SCALE) method to convert (19) to a convex problem. For more details and proofs, refer to [10].

Based on this method, a lower bound for user rate is adopted. Thus, the convex form of the online power allocation sub problem is as follows:

$$\begin{aligned}
&\max_{\tilde{\boldsymbol{P}}} \sum_{m=1}^{M} \sum_{n=1}^{N} \sum_{f=1}^{F} b_f^m \rho_{f,n}^m \psi_{f,n}^m \left(\log_2 (r_{f,n}^m) + \chi_{f,n}^m\right), \\
&\text{s.t} \ \ (19 \text{ b to } 19 \text{ f}). \qquad (20)
\end{aligned}$$

where $\psi_{f,n}^{m} = \frac{\gamma_{n,f}^{m}\left(\tilde{\boldsymbol{P}}(t-1)\right)}{1+\gamma_{n,f}^{m}\left(\tilde{\boldsymbol{P}}(t-1)\right)}$, $\chi_{f,n}^{m} = r_{m,f}^{n}\left(\boldsymbol{P}(t-1)\right) - \psi_{f,n}^{m}\log\left(\gamma_{m,f}^{n}\left(\boldsymbol{P}(t-1)\right)\right)$, $\boldsymbol{P} = \exp(\tilde{\boldsymbol{P}})$ and the index $(t-1)$ shows the last step in the iterations.

The power allocation sub-problem is solved by the Lagrangian dual decomposition method. The Lagrangian function is as follows

$$L_2\left(\boldsymbol{\rho}, \tilde{\boldsymbol{P}}, \boldsymbol{\nu}\right) = \sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{f=1}^{F} b_f^m \rho_{f,n}^m \psi_{f,n}^m\left(\log_2\left(r_{f,n}^m\right) + \chi_{f,n}^m\right) + \sum_{f=1}^{F}\nu_{1,f}\left(P_f^{max} - \sum_{m=1}^{M}\sum_{n=1}^{N}\rho_{f,n}^m \tilde{p}_{f,n}^m\right) \tag{21}$$

By solving $\frac{\partial L_2}{\partial \tilde{\boldsymbol{P}}} = 0$, $\tilde{\boldsymbol{P}}^*$ and then $\boldsymbol{P}^*$ is obtained and similar to offline dual Lagrange solution, the related Lagrange multipliers $\boldsymbol{\nu}$ are calculated using the sub-gradient method as:

$$\nu_{1,f}^{l+1} = \left[\nu_{1,f}^{l} - s^l\left(P_f^{max} - \sum_{m=1}^{M}\sum_{n=1}^{N}\rho_{f,n}^m \tilde{p}_{f,n}^m\right)\right]^{+}. \tag{22}$$

Similar to the offline problem, the KKT conditions are satisfied for the online sub problem (20). It is convex problem; therefore, the first level of optimality conditions is attained. Also the convergence is guaranteed provided that s^l is chosen to be square summable [26].

2) ONLINE PASE- CELL ALLOCATION

According to the main optimization problem (7), NSR is selected as the target function besides other constraints including coverage and HSR conditions.

$$\text{Target: } \max_{b_f^m} \sum_{f=1}^{F}\sum_{m=1}^{M}\sum_{n=1}^{N} b_f^m R_{f,n}^m; \tag{23a}$$

$$\text{s.t: } b_f^m \in \{0,1\}; \quad \forall m, f \tag{23b}$$

$$\sum_{f=1}^{F} b_f^m \leq 1; \quad \forall m \tag{23c}$$

$$\mathcal{P}_c \geq \mathcal{P}_{min}; \quad \forall m \tag{23d}$$

$\boldsymbol{B}$ is obtained from solving (23) using the dual Lagrange algorithm. Regarding the complexity of this solution and to guarantee fairness, it is possible to obtain sub optimal solution by dividing (23a) into following M equations:

$$\text{Target: } \max_{\boldsymbol{B}} \{b_f^m[t]\,\bar{r}_f^m \boldsymbol{P}[t-1]\}. \tag{24}$$

Subject to the constraints of problem (23). $\bar{r}_f^m$ is $E_n\{r_{n,f}^m\}$ or average user rate in terms of the subcarriers.

Problem (24) can be easily solved by using the search algorithms.

3) ONLINE PHASE -- SUBCARRIER ALLOCATION

The aim of subcarrier allocation problem is to maximize NSR by considering the fixed values of $\boldsymbol{P}[t-1]$ obtained from the previous power allocation iteration and $\boldsymbol{B}$ which is obtained from cell allocation sub problem. Based on the method in [7], it is possible to decompose (7) into M sub-problems, related to each user and each BS in order to find the maximum user rate by obtaining the best $\boldsymbol{\rho}$.

$$\text{Target: } \max_{\boldsymbol{\rho}} \{b_f^m[t-1]\,\rho_{f,n}^m[t]\,r_{f,n}^m(\boldsymbol{P}[t-1]\}, \tag{25a}$$

$$\rho_{f,n}^m \in \{0,1\}; \quad \forall m, f, n \tag{25b}$$

$$\sum_{m=1}^{M}\rho_{f,n}^m; \quad \forall f, n \tag{25c}$$

Equations (24) and (25) are mixed integer linear problems (MILP), thus it can be solved easily by pattern search techniques such as brand-and-branch algorithm [33].

Discussion 1: The online phase is a general structure with the possibility of applying different types of solutions such as a general dual decomposition like optimal-iterative spectrum balancing (OSB or ISB) [7], [10], [32], [33], and iterative water filling (IWF), [6]. All these solutions satisfy the KKT conditions and provide suboptimal results. The OSB technique results an upper limit of optimal points and approaches the optimum solutions within the limits of large numbers of N. Each algorithm provides different acceptable results from the main criteria of complexity, fairness, and convergence time point of view [7], [10]. Here, successive convex approximation (SCA) dual Lagrange solution is carried out in power allocation sub-problem to analyze the performance of the H-CRA and HP-CRA.

Discussion 2: As another solution to cell and resource allocation sub problems, it is possible to deploy joint solution of these sub-problems by mesh algorithms [34]. NOMAD [35] is a proper tool to solve mixed integer linear and nonlinear problems near-optimally but its complexity is too high.

Discussion 3: The objective and constraint functions in both power allocation and cell-resource allocation problems are differentiable. If $\boldsymbol{P}^*$ and $\boldsymbol{\nu}^*$ are the primal and dual optimal points of the solutions to (20) then the KKT conditions will be satisfied with proper values of the thresholds related to the constraints. Regarding the convexity of the online power allocation problem, (20) and the KKT conditions, the first level optimal power solution is attained. Besides, the other online subproblems (cell and subcarrier allocation problems) are linear whose solutions are globally optimal. Therefore, based on ASM, the proposed algorithm is converged to a local solution [32].

Discussion 4: As mentioned in Section II, we assume perfect CSI and backhaul capacity. However, it is also possible to consider imperfect assumptions: For the imperfect CSI case, it can be firstly modeled by the summation of the estimated CSI and its estimation error which is usually assumed to be a Gaussian random variable. Then, the actual CSI is modeled with the imperfect CSI [36]. Accordingly, the proposed method is applicable for the imperfect CSI case.

For the backhaul limited capacity scenario, it is only needed to add a new sum access rate constraint to the optimization problem [10]. By following the same line of arguments of the unlimited case, we can solve the corresponding limited optimization problem.

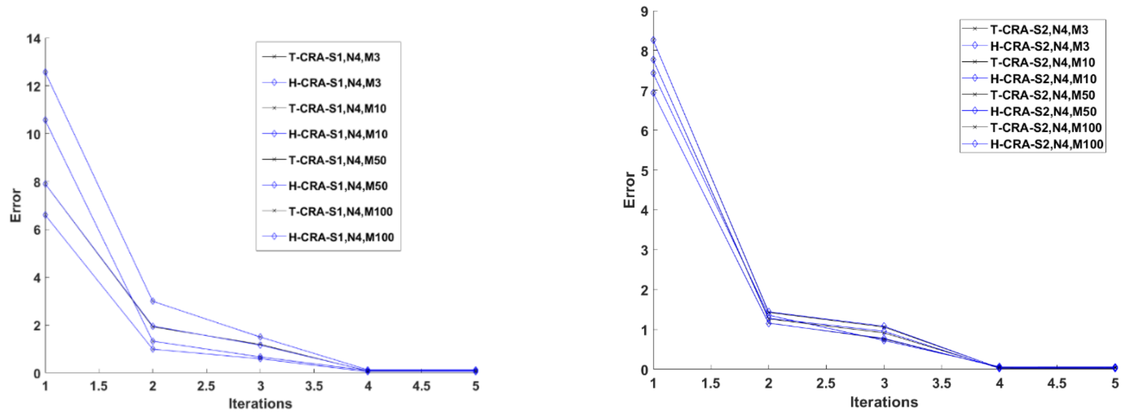


FIGURE 5. Convergence speed of H-CRA or HP-CRA and T-CRA for different scenarios: S1 and S2 and configurations of M and N.

4) BENCHMARKING ACROSS T-CRA AND HP-CRA

To justify the results of the proposed solution, the performance of HP-CRA and T-CRA are compared. HP-CRA is the same as H-CRA except the offline power allocation is removed and only $\boldsymbol{\theta}^*$ is obtained in the offline phase.

Under similar conditions of fairness, the same online algorithms of cell, power, and subcarrier allocation are utilized for T-CRA. Similar to the hybrid solution, these three sub-problems are solved iteratively following (16). The differences with H-CRA are: I) T-CRA is carried out only in online sense. II) T-CRA does not perform joint CRA and mobility analysis and the constrains of $\mathcal{P}_c$ and $\mathrm{E}_R\{HSR\}$ are not considered. The overall formulation is similar to the online part except the coverage probability constraint, and $\boldsymbol{P}^*_{Tier}$. Cell allocation is similar to the method of H-CRA.

In the hybrid method, the T-CRA power allocation sub-problem is solved by the dual decomposition method which obtains the upper bound of the optimal solutions. Also, cell-subcarrier allocation sub-problems are solved by the mentioned search method. The benchmarking is carried out by investigating the performances of the algorithms using different criteria such as convergence behavior, HSR or HFR, equivalent to (100-HSR), and NSR which presents spectrum efficiency.

V. NUMERICAL ANALYSIS

The performance of the proposed H-CRA approach is investigated using numerous simulations and comparison with the T-CRA and HP-CRA algorithms. The considered structure consists of joint cell-resource allocation and mobility. The mobility pattern is based on RWP Mobility as shown in Fig. 1. HO decision parameter is SINR. HO is performed after averaging over measured signals to take out the fast fading effects.

In addition, we consider three tiers (macro, micro and pico cells) with different $\boldsymbol{P}^{max} = [40, 2, 0.2]$ Watts and $\boldsymbol{\beta} = [1, 2, 3]$ to form HUDN. It is assumed that the spectrum is shared among all BSs and carrier frequency is 2.6 GHz. The noise power spectrum density is equal to -174dBm/Hz. The channel gain is considered to be the superposition of path loss, shadowing, and fast fading. In accordance with the dense urban environments, the path loss exponent, α, is 4. The fast-fading factor is based upon a normalized Rayleigh function for modeling the multi-path effect [37]–[39].

$\mathcal{A}$ is the domain of the network as a square-shaped cluster with the dimensions of 1000 m $\times$ 1000 m. In order to analyze the performance, the Monte Carlo method and the stochastic geometry model are adopted, hence, instead of the number of BSs their density (λ_i) are used and the number of BSs in each tier is determined based on PPP and λ_i. In each run of the simulation, the number and locations of BSs and users are changed.

A. CONVERGENCE BEHAVIOR

In order to study the behavior of the algorithms under different network's load, different numbers of users, M, and subcarriers, N, are evaluated. If the threshold of convergence is 0.1 and E_l is the error function in terms of $\boldsymbol{P}$ in the l^{th} iteration, then the iterations stop if $\|E_l - E_{l-1}\| \leq 0.1$ or the maximum number of iterations ($l_{max} = 500$) is reached. The typical convergence behavior of the different algorithms is presented in Fig. 5, which indicates the number of iterations up to convergence for different scenarios of M and N. Two settings for BSs density are chosen called S1 and S2 with $\boldsymbol{\lambda}_1 = [2e^{-3}, 3e^{-3}, 4e^{-3}]$ and$\boldsymbol{\lambda}_2 = \left[1.5e^{-3}, 2.25e^{-3}, 3e^{-3}\right]$, respectively. It is clear that the convergence speed is proper and similar to T-CRA methods.

B. MOBILITY MANAGEMENT PERFORMANCE

Numerical analysis is performed for various network configurations of $\boldsymbol{\lambda}$, M, and N. Note that M and $\frac{M}{N}$ are selected as network load metrics to analyze the performance. When the load or traffic is high, there is high probability of blockage or drop rate during HO process.

In addition, the higher load of the HUDNs raises the higher interference hence, NSR is saturated and HO failures increase due to the lower SINR in the cells' border.

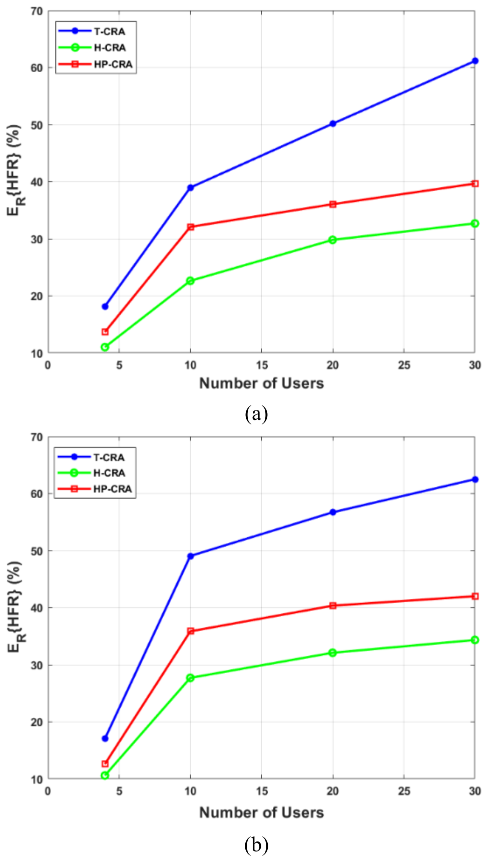


FIGURE 6. **(a) $E_R\{HFR\}$ versus the number of users, $N = 10$ and $\lambda_1 = \left[2e^{-3}, 3e^{-3}, 4e^{-3}\right]$. (b) $E_R\{HFR\}$ versus the number of users $N = 10$ and $\lambda_2 = [1 \cdot 5e^{-3}, 2 \cdot 25e^{-3}, 3e^{-3}]$.**

Fig. 6 and Fig. 7 compare the results of the handover failure rate $\mathrm{E}_R\{HFR\} = 100 - \mathrm{E}_R\{HSR\}$ of the three proposed solutions in terms of different values of $\boldsymbol{\lambda}$ and N.

Fig. 6a and Fig. 6b show the comparisons for two density sets: $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ when N =10, $\mathcal{P}_{min} = 0.9$ and $\tau = 0.9$. As seen in these figures, H-CRA and HP-CRA algorithms outperform the T-CRA method for different scenarios in all traffic conditions. For instance, Fig. 6 shows about 78% improvement of H-CRA in comparison with T-CRA when $N = 10$ and $M = 10$.

Moreover, H-CRA and HP-CRA are much more efficient in higher network's load compared to T-CRA. Furthermore, H-CRA is superior than HP-CRA about 20 % denoting the efficiency of offline power allocation. In fact, this improvement comes from considering jointly mobility pattern and channel information in the resource allocation procedure. In other words, unlike traditional works that is only the function of channel information, our methods are function of both channel and mobility information.

C. NETWORK SUM RATE

Here, the NSR of the algorithms, the sensitivity of $\mathcal{P}_{min}$, and τ are investigated.

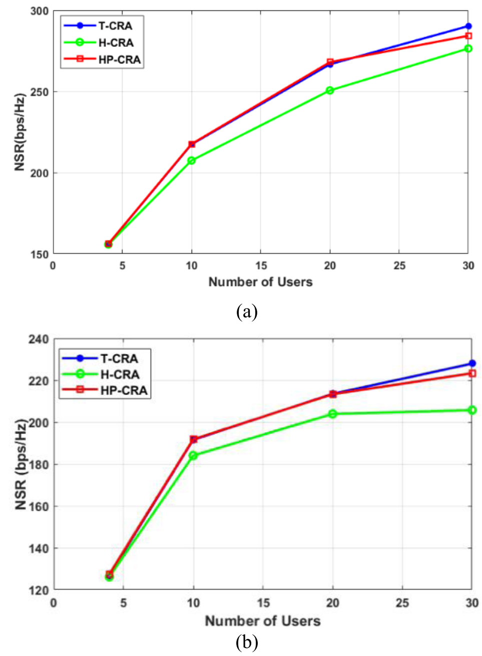


FIGURE 7. **(a) NSR versus the number of users, $N = 10$ and $\lambda_1 = [2e^{-3}, 3e^{-3}, 4e^{-3}]$. (b) NSR versus the number of users, $N = 10$ and $\lambda_2 = [1 \cdot 5e^{-3}, 2 \cdot 25e^{-3}, 3e^{-3}]$.**

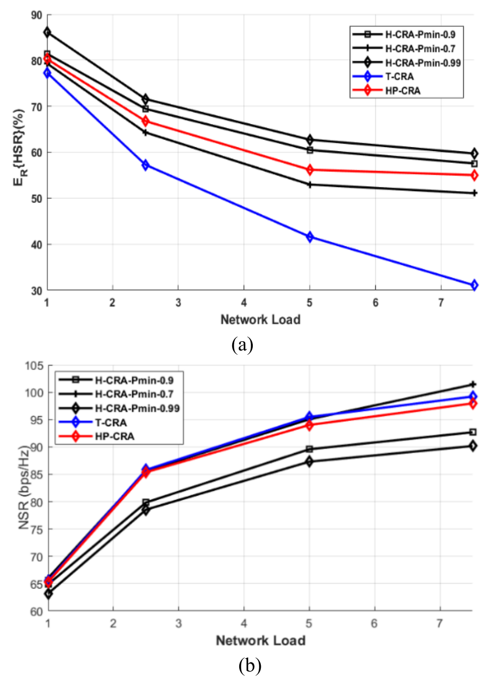


FIGURE 8. **(a) $E_R\{HSR\}$ versus the network load for the proposed algorithms and different values of $\mathcal{P}_{min}$. $N = 4$ and $\lambda = [2e^{-3}, 3e^{-3}, 4e^{-3}]$. (b) NSR versus the network load for the proposed algorithms and different values of $\mathcal{P}_{min}$. $N = 4$ and $\lambda = [2e^{-3}, 3e^{-3}, 4e^{-3}]$.**

Fig. 7 presents NSR with $\mathcal{P}_{min} = 0.9$ according to the two scenarios with $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ versus the network load. The results of simulations justify that the NSR of both H-CRA and HP-CRA are very similar and near to optimal values especially HP-CRA. Since the coverage and HSR constraints are added in H-CRA, and some parts of bandwidth are allocated to the mobile users, the NSR is reduced. However, the difference is less than 10% which is acceptable in practice.

$\mathcal{P}_{min}$ is an important parameter which has a key role in the performance where it makes a tradeoff between NSR and $\mathrm{E}_R\{HSR\}$. In fact, higher values of $\mathcal{P}_{min}$ results in higher $\mathrm{E}_R\{HSR\}$ and lower NSR. Therefore, the optimum value of $\mathcal{P}_{min}$ should be chosen based on the network configurations.

Fig. 8 introduces the effect of $\mathcal{P}_{min}$ with different values: $\{0.7, 0.9, 0.99\}$ on NSR and $\mathrm{E}_R\{HSR\}$. As it is seen, $\mathcal{P}_{min} = 0.9$ is a proper threshold value which leads to near-optimal values of $\mathrm{E}_R\{HSR\}$ and NSR. Note that the network load is considered as $\frac{M}{N}$.

VI. CONCLUSIONS

Heterogeneous ultra-dense networks' behavior has been investigated in terms of average handover success rate (HSR) and coverage probability through mathematical and numerical analysis stating that average HSR in these networks is equivalent to coverage probability. Furthermore, a novel mobility-aware cell-resource allocation structure is proposed. In this regard, we aim to maximize network sum rate (NSR) subject to the coverage and average HSR ($\mathrm{E}_R\{HSR\}$) constraints, transmit power limitation, and subcarrier and cell allocation restrictions. By dividing the main problem into ergodic and instantaneous sub-problems, two practical solutions, namely hybrid cell resource allocation (H-CRA) and hybrid partial cell resource allocation (HP-CRA) have been provided. This approach is flexible to the dynamic changes of the network including, topology, traffic, and mobility pattern. It has been demonstrated that $\mathrm{E}_R\{HSR\}$ results of both solutions outperform traditional cell resource allocation (T-CRA). Especially, H-CRA has remarkable improvement in average HSR about 78% compared to T-CRA as shown in Fig. 6 and Fig. 8.

TABLE 4. List of abbreviations.

HUDN	Heterogeneous ultra-dense network
RA	resource allocation
($\mathrm{E}_R\{HSR\}$)	expectation of Handover Success Rate
H-CRA	hybrid cell-resource allocation
HP-CRA	hybrid-partial CRA
T-CRA	traditional CRA
IoT	internet of things
SINR	signal to interference plus noise ratio
CRA	cell resource allocation
BSs	base stations
NSR	network sum rate
IRA	instantaneous resource allocation
ERA	ergodic resource allocation
CSI	channel state information
CDI	channel distribution information
RWP	Random waypoint
HSR	handover success rate
HFR	handover failure rate
HO	handover
RSS	received signal strength
OFDMA	Orthogonal Frequency Multiple Access
PPP	Poisson Point Process
PDF	probability distribution function
KKT	Karush_Kuhn_Tucker
SCALE	successive convex approximation with low complexity
MILP	mixed integer linear programming
OSB or ISB	optimal-iterative spectrum balancing
IWF	iterative water filling

APPENDIX

Proof of Lemma 1: Let the mobile user moves from point $\boldsymbol{r}_1$ to $\boldsymbol{r}_2$ in convex domain $\mathcal{A}$, facing with handover failure in $\boldsymbol{r}_2$. Considering that $\gamma^m_{f_i,n}(\boldsymbol{r}_2)$ is SINR related to connection of user m to cell f_i from tier i on subcarrier n also $\gamma^m_{g_i,k}(\boldsymbol{r}_1)$ is SINR at location $\boldsymbol{r}_1$ connecting user m to cell g_i belonging to tier i on subcarrier k. Thus, the probability of HFR at $\boldsymbol{r}_2$ is defined as:

$$\mathrm{P_{HFR}}(\boldsymbol{r}_2) = \mathbb{P}\left(\cap_{i=1}^{I}\gamma^m_{f_i,n}(\boldsymbol{r}_2) < \beta_i \mid \bigcup_{i=1}^{I}\gamma^m_{g_i,k}(\boldsymbol{r}_1) \geq \beta_i\right) \tag{a1}$$

$$= \mathbb{P}\left(\cap_{i=1}^{I}\gamma^m_{f_i,n}(\boldsymbol{r}_2) < \beta_i\right). \tag{a2}$$

Equation (a2) follows from the fact that $\gamma^m_{f,n}(\boldsymbol{r}_2)$ is independent of $\gamma^m_{g,k}(\boldsymbol{r}_1)$. Based on 3GPP standards, HO triggers after making L1 and L3 filtering on measurements in order to remove fast fading effects. Therefore, when the distance between these two points are more than typical thresholds, handover failure is independent from the last location situation. Thus, (a2) is equivalent to:

$$\mathrm{P_{HFR}}(\boldsymbol{r}_2) = 1 - \mathcal{P}_c(\boldsymbol{r}_2), \tag{a3}$$

where $\mathcal{P}_c(\boldsymbol{r}_2)$ follows from (4) or (5). Let the user faces with low quality at location dA and cannot handover to a better cell and taking into account the (4) and (a3). It is evident that $\mathcal{P}_c$ then $\mathrm{P_{HFR}}$ are independent of location. Thus, if the outage is occurred in a small area, dA, it follows $d(HFR) = P_{outage}dA$ and $P_{Outage} = 1 - \mathcal{P}_c$.

Therefore, the expectation over HFR in $\mathcal{A}$ is:

$$\mathrm{E}_R\{HFR\} = \int_A (1 - \mathcal{P}_c) f_R(\boldsymbol{r})\, d\boldsymbol{r}, \tag{a4}$$

where $f_R(\boldsymbol{r})$ is the pdf of mobile users' locations or

$$\mathrm{E}_R\{HSR\} = \mathcal{P}_c \int_A f_R(\boldsymbol{r})\, d\boldsymbol{r}. \tag{a5}$$

In both cases of uniform and non-uniform of distributions of users movement $f_R(\boldsymbol{r})$, (a5) is simplified as follows

$$\mathrm{E}_R\{HSR\} = \mathcal{P}_c. \tag{a6}$$

REFERENCES

[1] A. Nordrum. Popular Internet of Things Forecast of 50 Billion Devices by 2020 Is Outdated. IEEE Spectrum. Accessed: Aug. 2016. [Online]. Available: https://spectrum.ieee.org/tech-talk/telecom/internet/popular-internet-of-things-forecast-of-50-billion-devices-by-2020-is-outdated

[2] R. El Hattachi and J. Erfanian. (Feb. 17, 2015). *NGMN 5G White Paper, NGMN*. [Online]. Available: https://www.ngmn.org

[3] R. Q. Hu and Y. Qian, "An energy efficient and spectrum efficient wireless heterogeneous network framework for 5G systems," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 94–101, May 2014.

[4] M. Lauridsen, L. C. Gimenez, I. Rodriguez, T. B. Sorensen, and P. Mogensen, "From LTE to 5G for connected mobility," *IEEE Commun. Mag.*, vol. 55, no. 3, pp. 156–162, Mar. 2017.

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

[6] L. Venturino, N. Prasad, and X. Wang, "Coordinated scheduling and power allocation in downlink multicell OFDMA networks," *IEEE Trans. Veh. Technol.*, vol. 58, no. 6, pp. 2835–2848, Jun. 2009.

[7] Z. Shen, J. G. Andrews, and B. L. Evans, "Adaptive resource allocation in multiuser OFDM systems with proportional rate constraints," *IEEE Trans. Wireless Commun.*, vol. 4, no. 6, pp. 2726–2737, Nov. 2005.

[8] A. Sharifian, R. Schoenen, and H. Yanikomeroglu, "Joint realtime and nonrealtime flows packet scheduling and resource block allocation in wireless OFDMA networks," *IEEE Trans. Veh. Tech.*, vol. 65, no. 4, pp. 2589–2607, Apr. 2015.

[9] N. Mokari, H. Saeedi, and P. Azmi, "Quantized ergodic radio resource allocation in cognitive femto networks with controlled collision and power outage probabilities," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 11, pp. 2090–2104, Nov. 2014.

[10] M. Farokhi and A. Zolghadrasli, "Dynamic spectrum, subcarrier and power allocation in heterogeneous cellular networks based on interference and small cells' backhaul limitations," *Arabian J. Sci. Eng.*, vol. 42, no. 7, pp. 2685–2696, Sep. 2016, doi: 10.1007/s13369-016-2314-0.

[11] L. Musavian and Q. Ni, "Effective capacity maximization with statistical delay and effective energy efficiency requirements," *IEEE Trans. Wireless Commun.*, vol. 14, no. 7, pp. 3824–3835, Jul. 2015.

[12] A. Hanggoro and R. F. Sari, "Performance evaluation of the manhattan mobility model in vehicular ad-hoc networks for high mobility vehicle," in *Proc. IEEE Int. Conf. COMNETSAT*, Dec. 2013, pp. 31–36, doi: 10.1109/COMNETSAT.2013.6870855.

[13] X. Lin, R. K. Ganti, P. J. Fleming, and J. G. Andrews, "Towards understanding the fundamentals of mobility in cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 4, pp. 1686–1698, Apr. 2013.

[14] P. S. Prasad and P. Agrawal, "A generic framework for mobility prediction and resource utilization in wireless networks," in *Proc. 2nd IEEE COMSNETS Conf.*, Bangalore, India, Jan. 2010, pp. 1–10.

[15] V. Capdevielle, A. Feki, and E. Sorsy, "Joint interference management and handover optimization in LTE small cells network," in *Proc. IEEE ICC Conf.*, Jun. 2012, pp. 6769–6773.

[16] D. Lopez-Perez, I. Guvenc, and X. Chu, "Mobility management challenges in 3GPP heterogeneous networks," *IEEE Commun. Mag.*, vol. 50, no. 12, pp. 70–78, Dec. 2012.

[17] Y. Zhang, H. Long, Y. Peng, A. V. Vasilakos, and W. Wang, "QoE and energy efficiency aware resource allocation for OFDM systems in group mobility environments," *Int. J Commun. Syst.*, vol. 27, no. 12, pp. 3526–3544, Dec. 2014.

[18] *Key Performance Indicators (KPI) for the Evolved Packet Core (EPC), Release 10*, document TS 32.455, 3GPP, 2011.

[19] Q. Liu and C. W. Chen, "Smart downlink scheduling for multimedia streaming over LTE networks with hard handoff," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 11, pp. 1815–1829, Nov. 2015.

[20] A. S. Cacciapuoti, "Mobility-aware user association for 5G mmWave networks," *IEEE Access*, vol. 5, pp. 21497–21507, 2017.

[21] H. S. Dhillon, R. K. Ganti, F. Baccelli, and J. G. Andrews, "Modeling and analysis of K-tier downlink heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 550–560, Apr. 2012.

[22] A. H. Sakr and E. Hossain, "Location-aware cross-tier coordinated multipoint transmission in two-tier cellular networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 11, pp. 6311–6325, Nov. 2012.

[23] H. Elsawy, E. Hossain, and M. Haenggi, "Stochastic geometry for modeling, analysis, and design of multi-tier and cognitive cellular wireless networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 3, pp. 996–1019, 3rd Quart., 2013, doi: 10.1109/SURV.2013.052213.00000.

[24] G. George, R. K. Mungara, A. Lozano, and M. Haenggi, "Ergodic spectral efficiency in MIMO cellular networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 2835–2849, May 2017, doi: 10.1109/TWC.2017.2668414.

[25] G. Giambene and V. A. Le, "Analysis of LTE-A heterogeneous networks with SIR-based cell association and stochastic geometry," *IEEE J. Commun. Netw.*, vol. 20, no. 2, pp. 129–143, Apr. 2018.

[26] M. Haenggi, J. G. Andrews, F. Baccelli, O. Dousse, and M. Franceschetti, "Stochastic geometry and random graphs for the analysis and design of wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 7, pp. 1029–1046, Sep. 2009, doi: 10.1109/JSAC.2009.090902.

[27] W. C. Y. Lee, *Wireless and Cellular Telecommunication*. New York, NY, USA: McGraw-Hill, 2006.

[28] Y. Zaki, *Future Mobile Communications: LTE Optimization and Mobile Network Virtualization* (Advanced Studies Mobile Research Center Bremen). Wiesbaden, Germany: Springer, 2012.

[29] E. Hyytia, P. Lassila, and J. Virtamo, "Spatial node distribution of the random waypoint mobility model with applications," *IEEE Trans. Mobile Comput.*, vol. 5, no. 6, pp. 680–694, Jun. 2006.

[30] W. Yu and R. Lui, "Dual methods for nonconvex spectrum optimization of multicarrier systems," *IEEE Trans. Commun.*, vol. 54, no. 7, pp. 1310–1322, Jul. 2006.

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

[32] M. Moltafet, R. Joda, N. Mokari, M. R. Sabagh, and M. Zorzi, "Joint access and fronthaul radio resource allocation in PD-NOMA based 5G networks enabling dual connectivity and CoMP," *IEEE Trans. Commun.*, to be published, doi: 10.1109/TCOMM.2018.2865766.

[33] D. T. Ngo, S. Khakurel, and T. Le-Ngoc, "Joint subchannel assignment and power allocation for OFDMA femtocell networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 1, pp. 342–355, Jan. 2014.

[34] C. Audet, S. Le Digabel, and C. Tribes, "Dynamic scaling in the mesh adaptive direct search algorithm for blackbox optimization," *Optim. Eng.*, vol. 17, no. 2, pp. 333–358, 2016.

[35] S. Le Digabel, "Algorithm 909: NOMAD: Nonlinear optimization with the MADS algorithm," *ACM Trans. Math. Softw.*, vol. 37, no. 4, pp. 44-1–44-15, Feb. 2011.

[36] N. Mokari, M. R. Javan, and K. Navaie, "Cross-layer resource allocation in OFDMA systems for heterogeneous traffic with imperfect CSI," *IEEE Trans. Veh. Technol.*, vol. 59, no. 2, pp. 1011–1017, Feb. 2010.

[37] S. Hara and R. Prasad, *Multicarrier Techniques for 4G Mobile Communications*. Norwood, MA, USA: Artech House, 2003.

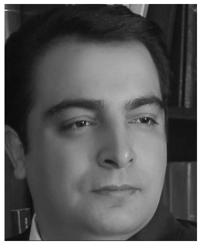
[38] T. S. Rappaport, *Wireless Communications: Principles and Practice*, vol. 2. Upper Saddle River, NJ, USA: Prentice-Hall, 1996.

[39] G. L. Stüber, *Principles of Mobile Communication*. Atlanta, GA, USA: Springer, 2017.

MASOUD FAROKHI received the B.S. degree from the University of Shiraz and the M.S. degree from the University of Tehran. He is currently pursuing the Ph.D. degree in telecommunication engineering with the University of Shiraz. He is also working in mobile network planning and optimization areas. His current research interests include 4G, 5G radio resource allocation, heterogeneous network, interference, and mobility management.

ALIREZA ZOLGHADRASLI was born in Fasa, Iran, in 1955. He received the B.Sc. degree (five years' program) in electrical and electronic engineering from Shiraz University (ancient Pahlavi University), Shiraz, Iran, in 1978, and the M.Sc. degrees in electronic, signal and data processing and the Ph.D. degree in image processing and communication systems from the Institute Polytechnic of Grenoble, Grenoble, France, in 1980, 1981, and 1985, respectively. He was an Assistant Professor and an Associate Professor with the University of Chambery, France, and the University of Grenoble, France, respectively, from 1985 to 1990.

From 1990 to 2001, he was an Assistant Professor with the Electrical Engineering Department, Shiraz University, where he was an Associate Professor from 2001 to 2016. He was promoted to Professor of communication systems at the Electrical Engineering Department in 2016. He has authored or co-authored more than 100 papers in international journals and conferences proceedings. His research interests include signal and image processing, communication, and wireless systems.



NADER MOKARI YAMCHI received the Ph.D. degree in electrical engineering from Tarbiat Modares University, Tehran, Iran, in 2014. In 2015, he joined the Department of Electrical and Computer Engineering, Tarbiat Modares University, as an Assistant Professor. He was involved in a number of large-scale network design and consulting projects in the telecom industry. His research interests include design, analysis, and optimization of communications networks.