Telecommunication Systems (2021) 77:753–766
https://doi.org/10.1007/s11235-021-00786-x

Power allocation and temporal fair user group scheduling for downlink NOMA

Eray Erturk[1] · Ozlem Yildiz[2] · Shahram Shahsavari[3] · Nail Akar[4]

Accepted: 1 April 2021 / Published online: 22 April 2021

Abstract

Non-Orthogonal Multiple Access (NOMA) has been proposed as a new radio access technique for cellular networks as an alternative to OMA (Orthogonal Multiple Access) in which the users of a group (pairs or triples of users in a group are considered in this paper) are allowed to use the wireless channel simultaneously. In this paper, for downlink single-input single-output SISO-NOMA, a heuristic power allocation algorithm within a group is first proposed which attempts to ensure that the users of a group benefit from simultaneous transmission equally in terms of achievable throughput. Moreover, a user group scheduling algorithm is proposed for downlink NOMA systems by which a user group is to be dynamically selected for transmission while satisfying long term temporal fairness among the individual contending users. The effectiveness of the proposed power allocation method along with the temporal fair scheduling algorithm for downlink NOMA is validated with simulations and the performance impact of the transmit power and the coverage radius of the base station as well as the number of users are thoroughly studied.

Keywords NOMA · Power allocation · User grouping · User group scheduling · Temporal fairness

1 Introduction

The advent of new technologies and applications such as virtual reality, augmented reality, and ultra high definition video streaming has led to a substantial increase in the demand for high throughput in cellular networks (see [1,6,7,9,22,28]). One of the fundamental approaches to increase the data rates in cellular networks is increasing spectral efficiency, i.e., the number of bits that can be transmitted per second in each unit of spectrum [6]. One of the proposed techniques to increase the spectral efficiency in cellular networks is non-orthogonal multiple access (NOMA) by which a BS can activate a user group (consisting of multiple users) at the same time-frequency resource block (RB) of the system either in the uplink (UL) or the downlink (DL) direction (for example see [11,12,18,25]). There are two main types of NOMA, namely power-domain NOMA and code-domain NOMA, that are described in detail in Islam et al. [11,14]. In this paper, we focus only on power-domain SISO-NOMA and refer to it as NOMA hereafter. While NOMA provides more transmission opportunities for users, it can potentially increase interference in the system due to multi-user activity. Consequently, proper resource allocation schemes including power allocation and user grouping and user group scheduling are key in coping with the additional interference and also in increasing the throughput. The survey paper by Islam et al. [14] presents advances regarding resource allocation focusing both on user pairing and power allocation in both SISO and multiple-input multiple-output (MIMO) settings. Another recent survey paper by Aldababsa et al. [3] provides

✉ Nail Akar
akar@ee.bilkent.edu.tr

Eray Erturk
erayerturk097@gmail.com

Ozlem Yildiz
zy2043@nyu.edu

Shahram Shahsavari
shahram.shahsavari.aut.ac@gmail.com

[1] Electrical and Computer Engineering, University of Southern California, Los Angeles, USA

[2] Electrical and Computer Engineering, NYU Tandon School of Engineering, New York University, New York, USA

[3] Electrical and Computer Engineering, University of Waterloo, Waterloo, Canada

[4] Electrical and Electronics Engineering, Bilkent University, 06800 Bilkent, Ankara, Turkey

a unified system model for NOMA including MIMO and cooperative communication scenarios.

We consider NOMA for downlink (DL) transmissions where a combination of superposition encoding and successive interference cancellation (SIC) decoding are used at the BS and the users, respectively (see [33]). To elaborate, each active user receives a superimposed signal including the signal of all active users. Successful SIC requires that each active user decodes the signals of the other active users with lower channel gains first, and then decodes its own signal by treating the remaining interfering signals as noise (see [39]). While the decoding procedure is in favor of users with higher channel gains as they can cancel more interfering signals, the power allocation is typically in favor of users with lower channel gains to compensate, i.e., the BS usually allocates more power to the signal of the users with weaker channel gains. The number of active users per RB is often limited by practical considerations such as SIC decoding complexity and latency which tend to increase with the number of multiplexed users (see [11]). Consequently, most of the prior works in the NOMA literature consider a small number of active users per RB such as two or three users (e.g., [2,23,27]). The performance of MIMO-NOMA is investigated in Zeng et al. [40] when multiple users are grouped into a cluster and the superiority of MIMO-NOMA over MIMO-OMA in terms of both sum channel capacity and ergodic sum capacity is proved analytically.

Furthermore, it is shown that in order to have higher throughput multiplexing gains with NOMA, it is essential to activate a set of users with sufficiently different channel gains (see [4]). As a result, given the propagation channel gains, several user grouping (a.k.a. user clustering or user pairing) approaches have been proposed in the literature to achieve this goal (e.g., [4,26]). Additionally, various user scheduling and power allocation algorithms have been proposed for NOMA systems with diverse objectives such as minimizing the probability of outage (see [38]), maximizing the system utility (e.g., [24]), or minimizing the power consumption (see [17]). Ideally, user group scheduling and power allocation should be optimized jointly, for example see [13]. However, such a joint optimization may be computationally prohibitive given the fact that computational power at the BS is limited. Consequently, it is more appealing to consider these two problems separately from a practical perspective, which is the approach we take in this paper.

Given a group of active users, power allocation with various objectives has been studied in the literature. For instance, the reference [23] proposes a power allocation for a DL NOMA system that secures higher data rates for the users of that group compared to an OMA system with round-robin scheduling. A related method is proposed by Cui et al. in [10] who study the problem of transmit power minimization with outage constraints for a given set of active users. In this paper, for the DL of a single-cell, single-carrier, time-slotted NOMA system, as our first contribution, we propose a heuristic low-complexity NOMA power allocation scheme on the basis of the results of Oviedo and Sadjadpour [23] for a given user group of pairs or triples while aiming for a balanced NOMA gain distribution across the users of the particular group.

The scheduler is in charge of choosing a group of users from a set of user groups for simultaneous transmission. Constructing the set of user groups is referred to as user grouping. Ali et al. [4] present an exhaustive list of user grouping schemes for NOMA. If all possible user groups are to be considered, then the complexity of the user group scheduler would be prohibitively high. Therefore, it is crucial to form user group sets with reduced cardinality so as to reduce the complexity of the scheduler. The second contribution of this paper is to evaluate (using simulations) the performance of several user grouping schemes with different complexities when a user group scheduler is in action.

In a practical DL NOMA system, given a user grouping and a power allocation algorithm, the allocated transmit power to each user of the candidate user groups are first calculated. Next, a scheduling algorithm selects one user group for each RB. Moreover, to achieve fairness, similar to OMA, the scheduling algorithms are often devised to maximize the system utility while satisfying a certain notion of fairness among users. Such schedulers are called opportunistically fair schedulers and have been studied extensively for OMA systems (e.g., see [21,29]). Indeed, several types of fair scheduling have been proposed in the literature including temporal fair (TF) [20], proportional fair (PF) [16], and utilitarian fair (UF) scheduling. While UF scheduling provides a guarantee for the users' utilities (e.g., throughput) (see [21,28]), TF scheduling ensures a fair temporal share (a.k.a. airtime share) allocation among the users (see [20,31,32]). Additionally, in PF scheduling, the objective is to maximize the sum of the logarithm of users' throughputs. The concavity of the log function provides a fair throughput allocation among the users (see [8,19]). The focus of this paper is on TF scheduling which can be more appealing for delay sensitive applications where the latency is of high importance (see [33,34]). While there is a vast literature on opportunistic TF scheduling in OMA systems (e.g., [20,21,29,30,37]), there are a limited number of prior works on TF scheduling for NOMA systems potentially due to the problem complexity. Given a power allocation scheme, the reference [33] proposes an optimal TF scheduler for NOMA systems in terms of system throughput under long-term minimum and maximum temporal share guarantees for each user while focusing on user pairs only. Also, optimal NOMA scheduling with short-term temporal fairness constraints is studied in Shahsavari et al. [34]. As our third contribution, we propose an opportunistic TF user group scheduler which attempts to maximize the system throughput while guaranteeing the same long-term temporal share for each user in the cell. We note that the structure of the proposed scheduler is inspired by its OMA counterpart presented in Shahsavari and Akar [29] and is also similar to the opportunistic TF NOMA scheduler suggested in Shahsavari et al. [33]. However, unlike [33] where minimum and maximum temporal share guarantees are considered for each user, our fairness constraint is to provide the same long-term temporal share for each user in the cell. Additionally, it can be shown that given our user grouping algorithm, our specific type of temporal fairness eliminates the chance of infeasibility of fairness constraints that may occur in the general scheme proposed by Shahsavari et al. [33]. Furthermore, Shahsavari et al. [33] focuses on optimal TF multi-user scheduling and considers a max-min power allocation in its numerical evaluations. However, we consider TF NOMA scheduling with a novel power allocation scheme and validate its effectiveness in terms of user and system throughput in our numerical evaluations. Also, our numerical results reveal that our proposed user scheduling and power allocation algorithm consistently outperforms an OMA system using an opportunistic scheduler in terms of both user and system throughput.

The rest of the paper is organized as follows. In Sect. 2, we describe the system model. The focus of Sect. 3 is on the proposed NOMA power allocation scheme elaborating on our proposed power allocation method for pairs and triples. In Sect. 4, we first describe the various user grouping mechanisms we study in this paper. Subsequently, we describe our proposed TF NOMA scheduler. In Sect. 5, we provide extensive numerical experimentation to evaluate the performance of the system and investigate the impact of user grouping and several system parameters on the performance. We conclude in Sect. 6.

2 System model

We consider the DL of a time-slotted single-cell single-frequency (or single-carrier) system consisting of a single-antenna BS as well as M single-antenna users. The time slots are indexed by $\tau \in \mathcal{Z}^{+}$, and the duration of each slot, T, is chosen to be less than the coherence time of the channel. Most of the well-established OMA schedulers for wireless cellular systems, such as the proportional-fair scheduler [16], temporal-fair scheduler [20], or the MaxWeight scheduler [35] are first proposed in the context of a single carrier system and these single-carrier schedulers are then appropriately extended to operate in multi-carrier settings. A straightforward extension method would be to schedule carriers one by one by using the single-carrier algorithm but more sophisticated methods targeting joint scheduling of carriers are also available; see for example several variations of the MaxWeight scheduling algorithm proposed by Andrews and Zhang [5] for a multi-carrier cellular network. While we consider single-carrier scheduling in this paper, we believe that our algorithms can be extended to the general multi-carrier scenarios such as Orthogonal Frequency Division Multiplexing (OFDM) systems. However, this line of research is left outside the scope of this paper.

We now describe the transmission model in the single-carrier OMA and NOMA systems with the following unifying treatment. Let a k-tuple of users with $1 \leq k \leq M$ are designated to be served in a given time slot τ. In the fair OMA system described in Oviedo and Sadjadpour [23], k users can be served in slot τ but one user at a time using Time Division Multiple Access (TDMA). In fair OMA, a single user is served for a duration of $\frac{T}{k}$ ensuring that each user in the designated k-tuple gets the same airtime within a time slot. On the other hand, in power-domain NOMA, the BS can transmit to these k users simultaneously. Let the users in the designated k-tuple be indexed as $i, i = 1, 2, \ldots, k$ and the complex channel coefficient for each user i at slot τ is denoted by $h_i(\tau)$, $i = 1, 2, \ldots, k$, with the corresponding channel gain $|h_i(\tau)|^2$, the latter being general continuous random variables with no a-priori assumptions imposed. We assume that the BS has a transmit power P. In the downlink of a NOMA system, the total transmit power P should be divided among the k users in the same k-tuple (see [14]). Let the power allocation coefficients for each user i be denoted by $p_i, i = 1, 2, \ldots, k$, satisfying $\sum_{i=1}^{k} p_i = 1$. Consequently, the transmit power allocated to user i is $p_i P$. Without loss of generality, we assume that the users are ordered such that $|h_1(\tau)|^2 < \ldots < |h_k(\tau)|^2$ which means each user i can perform SIC (Successive Interference Cancellation) at the receiver and remove the interference from the weaker users, but stronger users signals remain as interference; see [33,36]. The received signal at user i is written as (see [3])

$$y_i = h_i \sum_{j=1}^{k} \sqrt{p_j P}\, x_j + n_i, \tag{1}$$

where x_i is the information of user i with unit energy and n_i is zero mean complex Gaussian noise with variance σ^2. We also let $\xi = P/\sigma^2$ to represent the transmit SNR (Signal to Noise Ratio). Based on these definitions, the achievable NOMA rate for each user i at time slot τ in units of bits/s/Hz, denoted by $r_i^{(k)}(\tau)$, can be written as follows: (see [3,23])

$$r_i^{(k)}(\tau) = \log_2 \left(1 \frac{p_i \xi |h_i(\tau)|^2}{(\sum_{j=i+1}^{k} p_j)\xi |h_i(\tau)|^2 + 1} \right), \tag{2}$$

for $i = 1, 2, \ldots, k-1$. Since user k is the strongest user in the designated k-tuple, it can cancel the interference completely and its rate denoted by $r_k^{(k)}(\tau)$ is given as

$$r_k^{(k)}(\tau) = \log_2\left(1 + p_k \xi |h_k(\tau)|^2\right). \tag{3}$$

In a fair OMA system, the rate in units of bits/s/Hz for each user i in the designated k-tuple at slot τ, denoted by $\bar{r}_i^{(k)}(\tau)$ is given as: (see [36])

$$\bar{r}_i^{(k)}(\tau) = \frac{1}{k}\log_2\left(1 + \xi |h_i(\tau)|^2\right), \tag{4}$$

where the term $\frac{1}{k}$ emerges since the BS devotes all its transmission power to serve each user i for a duration of $\frac{T}{k}$; see [23]. Our proposed FAir NOMA Scheduler (FANS) works on a per-slot basis. At the beginning of each slot τ, we assume that the BS will find the power allocation coefficients for

(1) each pair of users from a pair set denoted by $\mathcal{P}(\tau)$, and
(2) each triple of users from a triple set denoted by $\mathcal{T}(\tau)$.

This problem is called the *power allocation* problem and the related algorithms we study in this paper are described in Sect. 3. On the other hand, the construction of the pair and triple sets $\mathcal{P}(\tau)$ and $\mathcal{T}(\tau)$ out of which a pair or a triple is selected, is called the *user grouping* problem and is key for the performance-complexity trade-off for the overall system. If these sets are chosen to be empty sets, then the system reduces to an OMA system with known performance setbacks. On the other hand, if these sets are chosen to be too wide, then the complexity of the scheduler will be high, prohibiting effective real-time operation. The user grouping algorithms that we study in this paper are described in Sect. 4. Subsequently, the BS will schedule a single user, or a user pair, or a user triple while satisfying long-term fairness constraints. This problem is known as the *user group scheduling* problem and the algorithm FANS that we propose for this purpose is presented in Sect. 4. To this end, we need several definitions for the temporal fairness attribute of the proposed scheduler FANS. For FANS, the long-term time utilization of all the users need to be the same to ensure temporal fairness among the users. To describe this requirement rigorously, we define $\rho_i(K)$, $0 \leq \rho_i(K) \leq 1, i = 1, 2, \ldots, M$ which is defined as the utilization factor of user i over a time span of K slots. Mathematically,

$$\rho_i(K) = \frac{\sum_{\tau=1}^{K} I\{\text{user } i \text{ is served at slot } \tau\}}{\sum_{\tau=1}^{K}\sum_{m=1}^{3} m\ I\{m \text{ users are simultaneously served at slot } \tau\}}, \tag{5}$$

where $I(\cdot)$ is the indicator function and is either 0 or 1 depending on whether the argument event is false or true, respectively. Mathematically, temporal fairness reduces to

$$\rho_i = \lim_{K\to\infty} \rho_i(K) = \frac{1}{M}, \quad 1 \leq i \leq M, \tag{6}$$

which ensures that the number of slots a single user is served in the long-term whether individually, or as part of a pair (or a triple), is the same across all users in the system. Since exact fulfillment of the condition in (6) is hard to achieve for finite K, in the numerical experiments, we will investigate a scalar fairness index known as the Jain's Fairness Index (JFI) [15]:

$$JFI(K) = \frac{\left(\sum_{i=1}^{M}\rho_i(K)\right)^2}{M\sum_{i=1}^{N}\rho_i(K)^2}, \quad \frac{1}{M} \leq JFI(K) \leq 1. \tag{7}$$

When $JFI(K)$ is close to unity, then we say the system is temporal fair over a span of K time slots. On the other hand, when $JFI(K)$ moves away from unity, then some users are said to be penalized in terms of temporal utilization of system resources.

We also define the average throughput of the user i in units of bits/s/Hz, denoted by $r_i(K)$, over a span of K slots as:

$$r_i(K) = \frac{1}{K}\sum_{\tau=1}^{K} r_i^{(m)}(\tau) I\{\text{user } i \text{ served, } m \text{ users scheduled, at slot } \tau\}, \tag{8}$$

for $m = 1, 2, 3$. The average system-wide throughput $r(K)$ over a span of K slots is then given by

$$r(K) = \sum_{i=1}^{M} r_i(K). \tag{9}$$

Finally, we define the steady-state throughput per user i, denoted by r_i, and also the overall system-wide throughput r as follows:

$$r_i = \lim_{K\to\infty} r_i(K), \ 1 \leq i \leq M, \quad r = \lim_{K\to\infty} r(K). \tag{10}$$

In addition to providing temporal fairness among users, when a pair or triple is selected by the scheduler, the fair power allocation to be described in the next section will be used to ensure fairness among users that make up the selected pair/triple. We note that unfair power allocations can give rise to unfair rate allocations despite the fulfillment of the temporal fairness condition given in (6).

3 Power allocation in NOMA

In Oviedo and Sadjadpour [23], the power allocation coefficients for NOMA are to be chosen to guarantee that for each user i in the k-tuple, the per-user rate achieved by

NOMA will be at least equal to the per-user rate achieved by OMA:

$$g_i^{(k)}(\tau) = \frac{r_i^{(k)}(\tau)}{\bar{r}_i^{(k)}(\tau)} \geq 1, \;\; 1 \leq i \leq k, \tag{11}$$

where $g_i^{(k)}(\tau)$ is the NOMA gain at slot τ achieved by user i in the k-tuple with respect to OMA. The condition (11) ensures that NOMA is advantageous to OMA not only for the sum rate of users but also for each user within the k-tuple. Oviedo and Sadjadpour [23] provide a means for obtaining the power allocation coefficients for a NOMA system satisfying (11) while naming the proposed methodology "fair NOMA". In this paper, we describe three power allocation methods on the basis of the work of Oviedo and Sadjadpour [23] for the two particular choices of $k = 2, 3$ with all the three methods satisfying the inequality (11) for all the users of the k-tuple. The case for $k > 3$ is left for future research.

3.1 Power allocation for pairs

In this section, the NOMA power allocation method proposed by Oviedo and Sadjadpour [23] for the case $k = 2$ is presented for the sake of completeness. In this case, we have two users 1 and 2 at slot τ with channel gains satisfying $|h_2(\tau)|^2 > |h_1(\tau)|^2$. The per-user rates $r_1^{(2)}(\tau)$ and $r_1^{(2)}(\tau)$ (in units of bits/s/Hz) are given in terms of the power allocation coefficients in (2) and (3). The sum rate, denoted by $r_{(1,2)}(\tau)$, for the ordered user pair (1, 2) is obtained by

$$r_{(1,2)}(\tau) = r_1^{(2)}(\tau) + r_2^{(2)}(\tau). \tag{12}$$

What now remains to be shown is how to find the power allocation coefficients p_1 and p_2 so that the identity (11) is satisfied. In order for $r_1^{(2)}(\tau) \geq \bar{r}_1^{(2)}(\tau)$ to hold true, the following inequality should be satisfied (see [23]):

$$p_2 \leq p_{2,max} = \frac{(1 + \xi|h_1(\tau)|^2)^{\frac{1}{2}} - 1}{\xi|h_1(\tau)|^2}. \tag{13}$$

On the other hand, the condition $r_2^{(2)}(\tau) \geq \bar{r}_2^{(2)}(\tau)$ yields a lower bound for p_2 (see [23]):

$$p_2 \geq p_{2,min} = \frac{(1 + \xi|h_2(\tau)|^2)^{\frac{1}{2}} - 1}{\xi|h_2(\tau)|^2}. \tag{14}$$

It can be shown as in Oviedo and Sadjadpour [23] that $p_{2,max} \geq p_{2,min}$ and a choice of $p_2 \in [p_{2,min}, p_{2,max}]$ and $p_1 = 1 - p_2$ is sufficient to ensure a NOMA allocation (11) which is superior to OMA for both users. Moreover, the policy of choosing $p_2 = p_{2,max}$ is shown in Oviedo and Sadjadpour [23] to maximize the sum rate $r_{(1,2)}(\tau)$ and is hence referred to as MSRPA (Maximum Sum Rate Power Allocation). However, with MSRPA, only the user 2 benefits from NOMA and user 1 does not benefit from NOMA at all. To address the NOMA gain asymmetry between the two users, we propose Fair Power Allocation (FPA) policy which ensures that both users benefit from the deployment of NOMA equally by the choice of p_2, i.e., the identity $g_1^{(2)}(\tau) = g_2^{(2)}(\tau)$ is forced to hold. FPA involves the solution of a nonlinear equation subject to the constraint $p_2 \in [p_{2,min}, p_{2,max}]$ to obtain p_2. For the purpose of achievement of balanced NOMA gains but with a lower complexity than FPA, we propose a heuristic power allocation method for which we choose p_2 to be the arithmetic average of $p_{2,min}$ and $p_{2,max}$, named HPA (Heuristic Power Allocation). The underlying motivation for HPA is to facilitate computation since the power allocation coefficients are to be dynamically computed for many user pairs at every time slot τ while choosing the best user pair among all the candidate pairs. Therefore, the computation needed for power allocation for each candidate pair needs to be kept at a minimum which is the main reason for proposing HPA as opposed to FPA.

In order to comparatively evaluate the three power allocation methods described above, we fix $\xi = 30$ dB, $|h_2(\tau)|^2 = 1$. Subsequently, the sum rate $r_{(1,2)}(\tau)$ and the NOMA gains $g_j^{(2)}(\tau)$, $j = 1, 2$ are plotted in Fig. 1 as a function of the varying channel gain $|h_1(\tau)|^2$ of user 1. We note that the power allocation coefficients for the particular case of FPA are obtained by exhaustive search. We observe that the performance of the proposed HPA policy is similar to that of FPA and thus provides a more balanced NOMA gain across the two users of the pair in comparison with MSRPA for which user 1 will always have a unit NOMA gain. Moreover, HPA slightly outperforms FPA in terms of the sum rate and the deviation of the performance of HPA from FPA increases with decreased channel gain $|h_1(\tau)|^2$ of user 1. We thus conclude that HPA is a low-computation alternative to FPA in the case of user pairs.

3.2 Power allocation for triples

In this subsection, $k = 3$ for which we have three users i, $i = 1, 2, 3$ at slot τ with channel gains satisfying $|h_3(\tau)|^2 > |h_2(\tau)|^2 > |h_1(\tau)|^2$. The per-user rates $r_i^{(3)}(\tau)$, $i = 1, 2, 3$ in units of bits/s/Hz are given in terms of the power allocation coefficients p_i, $i = 1, 2, 3$, $0 \leq p_i \leq 1$, $p_1 + p_2 + p_3 = 1$, as in (2) and (3). On the other hand, the sum rate, denoted by $r_{(1,2,3)}(\tau)$, for the ordered user triple (1, 2, 3) is simply the sum of the three per-user rates:

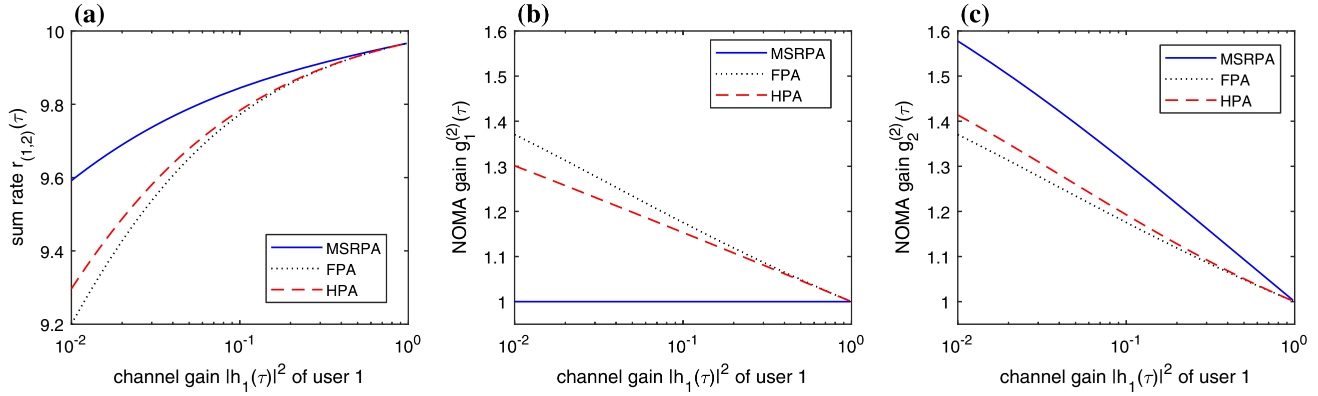


Fig. 1 **a** Sum rate $r_{(1,2)}(\tau)$, **b** NOMA gain $g_1^{(2)}(\tau)$, **c** NOMA gain $g_2^{(2)}(\tau)$ as a function of the channel gain $|h_1(\tau)|^2$ of user 1

$$r_{(1,2,3)}(\tau) = r_1^{(3)}(\tau) + r_2^{(3)}(\tau) + r_3^{(3)}(\tau). \tag{15}$$

For fair NOMA power allocation, the identity (11) should hold for $k = 3$. It is easy to show that the condition $r_3^{(3)}(\tau) \geq \bar{r}_3^{(3)}(\tau)$ yields the following lower bound for p_3:

$$p_3 \geq p_{3,min} = \frac{(1+\xi|h_3(\tau)|^2)^{\frac{1}{3}} - 1}{\xi|h_3(\tau)|^2}. \tag{16}$$

On the other hand, the condition $r_2^{(3)}(\tau) \geq \bar{r}_2^{(3)}(\tau)$ results in an upper bound for p_3 depending on p_2:

$$\begin{aligned} p_3 &\leq p_{3,max}(p_2) \\ &= \left(\frac{p_2\xi|h_2(\tau)|^2}{(1+\xi|h_2(\tau)|^2)^{1/3} - 1} - 1\right)\frac{1}{\xi|h_2(\tau)|^2}, \end{aligned} \tag{17}$$

and a lower bound for p_2 depending on p_3:

$$\begin{aligned} p_2 &\geq p_{2,min}(p_3) \\ &= \frac{\big((1+\xi|h_2(\tau)|^2)^{1/3} - 1\big)\big(p_3\xi|h_2(\tau)|^2 + 1\big)}{\xi|h_2(\tau)|^2}. \end{aligned} \tag{18}$$

Finally, the inequality $r_1^{(3)}(\tau) \geq \bar{r}_1^{(3)}(\tau)$ results in the following upper bound p_2 depending on p_3:

$$\begin{aligned} p_2 &\leq p_{2,max}(p_3) \\ &= \left(\frac{\xi|h_1(\tau)|^2 - \big((1+\xi|h_1(\tau)|^2)^{1/3} - 1\big)}{\xi|h_1(\tau)|^2(1+\xi|h_1(\tau)|^2)^{1/3})} - p_3\right). \end{aligned} \tag{19}$$

Similar to the case of pairs of users, we define the policy MSRPA that maximizes the sum rate subject to the condition (11) and the policy FPA by which users of the triple benefit from the use of NOMA equally in comparison with OMA. As a low-computation alternative, we propose HPA for triples of users by setting

$$p_i = \frac{1}{2}(p_{i,min} + p_{i,max}), i = 2, 3, \tag{20}$$

which results in a linear equation with two unknowns that are to be solved to obtain the power allocation coefficients p_2 and p_3, which is straightforward to write from the identities (16)–(19). Subsequently, p_1 is set to $1 - p_2 - p_3$. In order to comparatively evaluate the three power allocation methods described above, we fix $\xi = 30$ dB, $|h_3(\tau)|^2 = 1$, $|h_1(\tau)|^2 = 10^{-2}$. Subsequently, the sum rate $r_{(1,2,3)}(\tau)$ and the NOMA gains $g_j^{(3)}(\tau)$, $j = 1, 2, 3$, using these three policies are plotted in Fig. 2 as a function of the channel gain $|h_2(\tau)|^2$ of user 2. The power allocation coefficients for the particular case of FPA are obtained by exhaustive search as in the case of pairs. We observe that in MSRPA, the NOMA gains of users 1 and 2 are one and the power allocation coefficients can be obtained by the simultaneous solution of the following two linear equations

$$p_2 = p_{2,max}(p_3), \;\; p_3 = p_{3,max}(p_2) \tag{21}$$

and is simple to implement. However, MSRPA suffers from unbalanced NOMA gains across the three users and all the NOMA benefit is enjoyed by user 3 alone, while users 1 and 2 do not benefit from NOMA at all, in comparison with OMA. On the other hand, HPA provides a more balanced NOMA gain distribution while slightly lagging the MSRPA policy in terms of the sum rate. We also observe that part of the NOMA gain enjoyed by user 3 in MSRPA appears to be shared almost equally by users 1 and 2 in HPA for varying values of $|h_2(\tau)|^2$. Stemming from lower computational complexity than FPA and a more balanced NOMA gain distribution than MSRPA, we will employ HPA for pairs and triples, in the numerical examples of this paper.

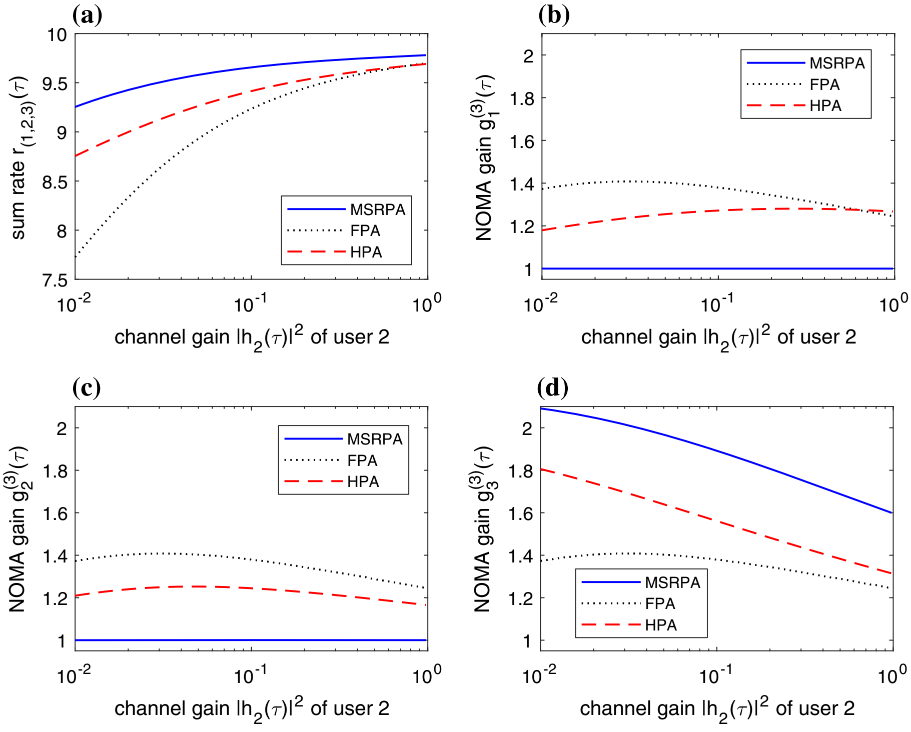


Fig. 2 **a** Sum rate $r_{(1,2,3)}(\tau)$, **b** NOMA gain $g_1^{(3)}(\tau)$, **c** NOMA gain $g_2^{(3)}(\tau)$, **d** NOMA gain $g_3^{(3)}(\tau)$ as a function of the channel gain $|h_2(\tau)|^2$ of user 2

4 User grouping and user group scheduling

4.1 User grouping

At the beginning of each slot τ, we assume that the BS will schedule the transmission to either a single user, a user pair, or a user triple, out of the set of given user groups, and then the HPA policy described in the previous section will be used during transmission for the selected pair or triple. For the description of the user grouping schemes of interest, we need the following definitions. Let $\mathcal{I}$ denote the set of individual users i, $1 \leq i \leq M$, and let $\mathcal{P}^{(1)}(\tau)$ denote the set of all candidate ordered user pairs (i, j) such that $1 \leq i, j \leq M$, $|h_j(\tau)|^2 > |h_i(\tau)|^2$. Similarly, let $\mathcal{T}^{(1)}(\tau)$ denote the set of all candidate ordered user triples (i, j, k) such that $1 \leq i, j, k \leq M$, $|h_k(\tau)|^2 > |h_j(\tau)|^2 > |h_i(\tau)|^2$. Above, candidacy is with respect to NOMA transmission. Note that the cardinalities of the sets $\mathcal{P}^{(1)}(\tau)$ and $\mathcal{T}^{(1)}(\tau)$ are $M(M-1)/2$ for $M \geq 2$ and $M(M-1)(M-2)/6$ for $M \geq 3$, respectively, making it computationally very expensive to dynamically select pairs and triples of users. In order to reduce the computational load of the user group scheduling algorithm, we propose to use certain user grouping schemes with lesser cardinalities than $\mathcal{P}^{(1)}(\tau)$ and $\mathcal{T}^{(1)}(\tau)$; see [4] for an exhaustive list of user grouping schemes for NOMA.

For the purpose of user grouping for pairs, we first rank the users according to their channel gains $|h_i(\tau)|$ so that rank of user i is denoted by $v_i(\tau)$. In this case, the user with the strongest channel has rank M, the user with the next strongest channel has rank $M-1$, and so on. When M is even, we propose to use the following subset of pairs $\mathcal{P}^{(2)}(\tau)$ for user grouping:

$$\mathcal{P}^{(2)}(\tau)=\left\{(i,j) : v_i(\tau)+v_j(\tau)=M+1,\, v_i(\tau)\leq v_j(\tau)\right\}$$
$$\bigcup\left\{(i,j) : v_j(\tau)-v_i(\tau)=\frac{M}{2},\, v_i(\tau)\leq v_j(\tau)\right\}. \tag{22}$$

With this construction, $|\mathcal{P}^{(2)}(\tau)| = M$ and all the individual users appear the same number of times, i.e., twice, in this proposed set. While trying to be fair among users during the set construction in terms of the number of occurrences, we also try to pair users with as much different channel gains as possible. When M is odd, we propose to add a dummy user with random channel gain and apply the procedure above but omit the pairs that contain the dummy source at the end.

For the purpose of user grouping for triples, when M is divisible by 6, we propose to use the following subset of triples $\mathcal{T}^{(2)}(\tau)$:

$$\mathcal{T}^{(2)}(\tau) = \left\{(i,j,k) : v_i(\tau) \leq \frac{M}{3},\, v_i(\tau)+v_k(\tau) = M+1,\right.$$
$$\left. v_j(\tau)-v_i(\tau) \text{ or } v_k(\tau)-v_j(\tau) = \frac{M}{3}\right\}. \tag{23}$$

With this construction, $|\mathcal{T}^{(2)}(\tau)| = \frac{2M}{3}$ and all the individual users appear twice in this subset. We also propose to use an alternative triple set $\mathcal{T}^{(3)}(\tau)$ with cardinality M as follows:

$$\mathcal{T}^{(3)}(\tau) = \mathcal{T}^{(2)}(\tau) \bigcup \Big\{(i, j, k) : v_j(\tau) - v_i(\tau) = v_k(\tau) - v_j(\tau) = \frac{M}{3}, v_i(\tau) \le v_j(\tau) \le v_k(\tau)\Big\}. \tag{24}$$

Moreover, similar to the pair set construction proposed above, we try to group users with as much different channel gains as possible while maintaining simplicity. When M is not divisible by 6, we propose to add up to 5 dummy users again with random channel gains to fulfill the divisibility-by-6 condition and apply the same procedure above. In the end, triples that contain one of these dummy sources are omitted if the remaining pair is already in the proposed set of pairs. Similarly, triples that contain two or more of these dummy sources are to be omitted as well.

4.2 User group scheduling

We are now ready to describe the proposed scheduler FANS that operates on a per slot basis. FANS maintains a per-user bucket variable, denoted by $b_i(\tau) \in \mathbb{R}$, for each user i at slot τ. Let $b(\tau)$ be the the row vector of per-user bucket values which is of size M. Also let e_i be a row vector of zeros of size M except for the ith position (which is set to one) and let e be a row vector of ones of size M.

The FANS scheduler operation at slot τ is described in detail in Algorithm 1. The bucket values are initially set to zero. In this algorithm, based on the channel gains of each user at slot τ, the subsets of pairs and triples that contain candidate pairs or triples to be served, are first constructed. In the most crucial step of the algorithm, one of the following will be scheduled for transmission:

- A user with the largest sum of rate and bucket values,
- A pair with the largest sum of sum rates and sum of bucket values,
- A triple with the largest sum of sum rates and sum of bucket values,

After the scheduling decision is made in favor of a user either individually or as part of a pair or a triple, this particular user's bucket value is incremented by α so that it will have a lesser chance at the forthcoming slots to be scheduled. Subsequently, all the bucket values are simultaneously decremented by a value in such a way that the sum of the bucket values remains as zero. The computational complexity of the algorithm is $\mathcal{O}(M)$ which involves finding the power allocation coefficients for each user group in the indicated group sets along with the sum rates of each group in line with the HPA algorithm proposed in Section 3. The storage requirement of the proposed algorithm is $\mathcal{O}(M)$ since only a bucket is maintained for each user holding a scalar real value.

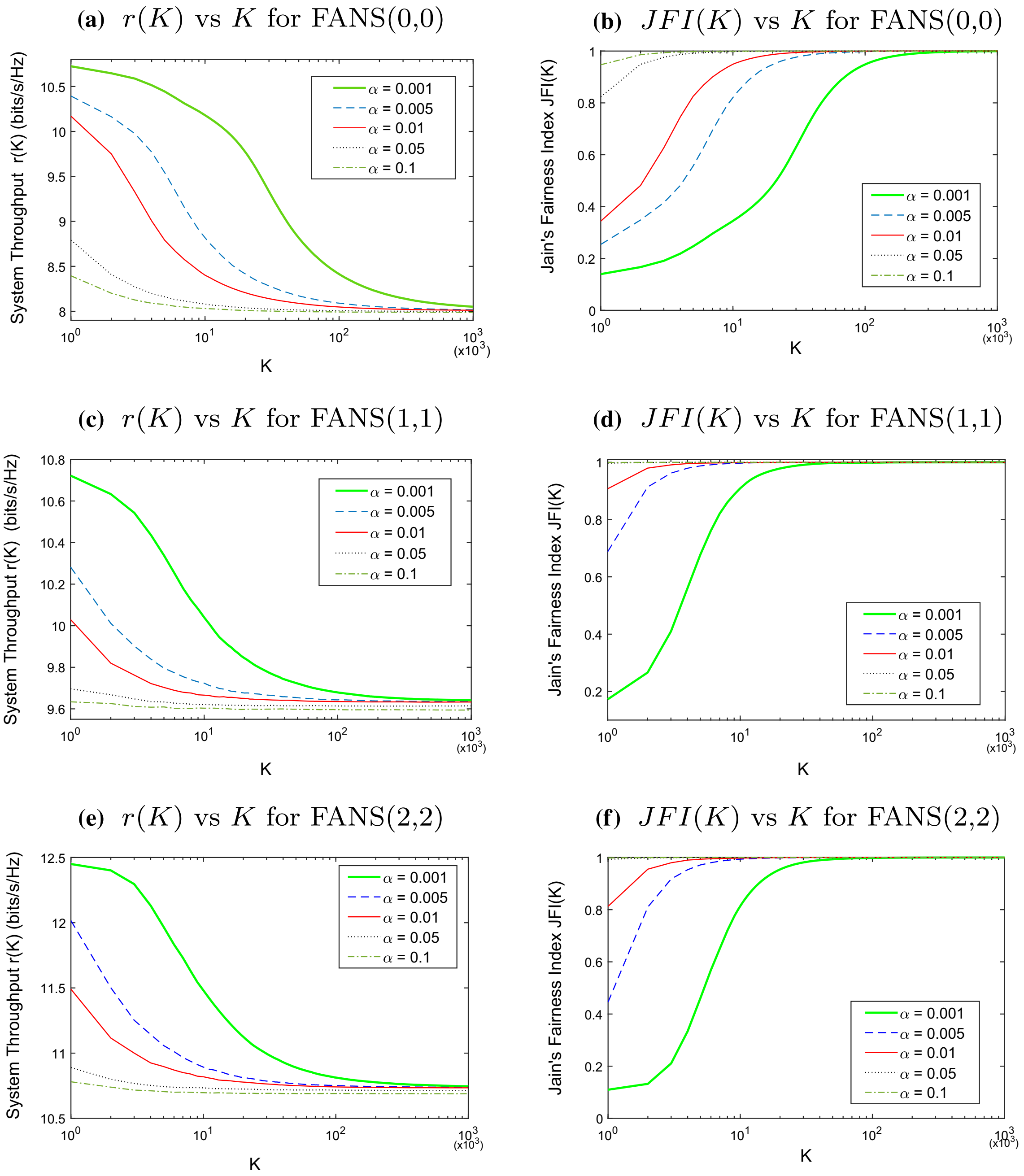
```
Data: time slot $\tau$, vector of channel gains $\{h_i(\tau)\}, i = 1, \ldots, M$,
      bucket vector $b(\tau)$, algorithm variable $\alpha$
Result: select a user, pair, or triple, for NOMA transmission and
        update the bucket vector
for each user $i$ do
    obtain $r_i^{(1)}(\tau)$ using $r_i^{(1)}(\tau) = \log_2(1 + \xi|h_i(\tau)|^2)$.
end
construct the set of pairs $\mathcal{P}(\tau)$ ($\emptyset$ or $\mathcal{P}^{(1)}(\tau)$ or $\mathcal{P}^{(2)}(\tau)$) ;
for each user pair $(m, n) \in \mathcal{P}(\tau)$ do
    obtain $r_{(m,n)}^{(2)}(\tau)$ using (12).
end
construct the set of triples $\mathcal{T}(\tau)$ ($\emptyset$ or $\mathcal{T}^{(1)}(\tau)$ or $\mathcal{T}^{(2)}(\tau)$ or
$\mathcal{T}^{(3)}(\tau)$ ) ;
for each user triple $(x, y, z) \in \mathcal{T}(\tau)$ do
    obtain $r_{(x,y,z)}^{(3)}(\tau)$ using (15).
end
for $i \in \mathcal{I}$, $(m, n) \in \mathcal{P}(\tau)$, $(x, y, z) \in \mathcal{T}(\tau)$, schedule the user $i^*$
or the user pair $(m^*, n^*)$, or the user triple $(x^*, y^*, z^*)$ that
maximize
$r_i^{(1)}(\tau) + b_i, r_{(m,n)}^{(2)}(\tau) + b_m + b_n, r_{(x,y,z)}^{(3)}(\tau) + b_x + b_y + b_z$ ;
if user $i^*$ is scheduled then
    $b(\tau + 1) = b(\tau) + \alpha(\frac{1}{M}e - e_{i^*})$;
else if user pair $(m^*, n^*)$ is scheduled then
    $b(\tau + 1) = b(\tau) + \alpha\left(\frac{2}{M}e - e_{m^*} - e_{n^*}\right)$;
else
    user triple $(x^*, y^*, z^*)$ is scheduled;
    $b(\tau + 1) = b(\tau) + \alpha\left(\frac{3}{M}e - e_{x^*} - e_{y^*} - e_{z^*}\right)$;
end
```

Algorithm 1: The pseudo-code for the proposed scheduler FANS at time slot τ.

Table 1 System parameters used in the numerical examples

Parameter	Value
Cell radius R (m)	100, 500, 1000
System bandwidth	10 MHz
System frequency	2.5 GHz
Path loss model	$128.1 + 37.6\log_{10}(\cdot)$ (in km)
Noise spectral density	−174 dBm/Hz
Shadowing standard deviation	8 dB
Number of users M	12, 18, 36, 72
BS transmit power budget P	30 dBm, 40 dBm

We have the following observations: The structure of the scheduler FANS for OMA systems, i.e., $\mathcal{P}(\tau) = \mathcal{T}(\tau) = \emptyset$, is known to be optimal in terms of the overall throughput satisfying temporal fairness: see Liu et al. [20] and Shahsavari and Akar [29]. In this paper, we extend the same structure to NOMA systems with dynamic k-tuple selection for $k \le 3$. For each choice of $\alpha > 0$, the proposed algorithm satisfies the fairness constraint (6) since the way the scheduler works, the individual bucket values can not diverge to minus or plus infinity. As long as the bucket values remain bounded for all K, one can easily show that (6) holds. However, the speed at which $\rho_i(K)$ approaches to $1/M$ in (6) depends on the partic-



(a) $r(K)$ vs K for FANS(0,0)

(b) $JFI(K)$ vs K for FANS(0,0)

(c) $r(K)$ vs K for FANS(1,1)

(d) $JFI(K)$ vs K for FANS(1,1)

(e) $r(K)$ vs K for FANS(2,2)

(f) $JFI(K)$ vs K for FANS(2,2)

Fig. 3 System-wide throughput $r(K)$ and Jain's fairness index $JFI(K)$ plotted as a function of K for five different values of the algorithm parameter α and for the three algorithm versions FANS(0,0), FANS(1,1), FANS(2,2)

ular choice of the parameter α. When α is relatively small, this convergence rate is lower but the overall system throughput is higher. On the other extreme, when α is large, the overall system throughput is much lower but the system becomes fair even for small values of K. We refer the reader to Shahsavari et al. [30] for more exhaustive analysis of the impact of the algorithm parameter α on convergence but in the context of multicell OMA systems. The proposed scheduler FANS has several variations depending on the choices of the sets $\mathcal{P}(\tau)$ and $\mathcal{T}(\tau)$. FANS(i, j) for $i = 0, 1, 2$ and $j = 0, 1, 2, 3$ uses $\mathcal{P}^{(i)}(\tau)$ for the set of pairs and $\mathcal{T}^{(j)}(\tau)$ for the set of triples where by convention $\mathcal{P}^{(0)}(\tau) = \mathcal{T}^{(0)}(\tau) = \emptyset$. In the numerical examples, we will compare and contrast the six user group scheduling algorithms FANS$(0, 0)$ (also referred to as OMA), FANS$(1, 0)$, FANS$(2, 0)$, FANS$(1, 1)$, FANS$(2, 2)$, and FANS$(2, 3)$.

5 Numerical results

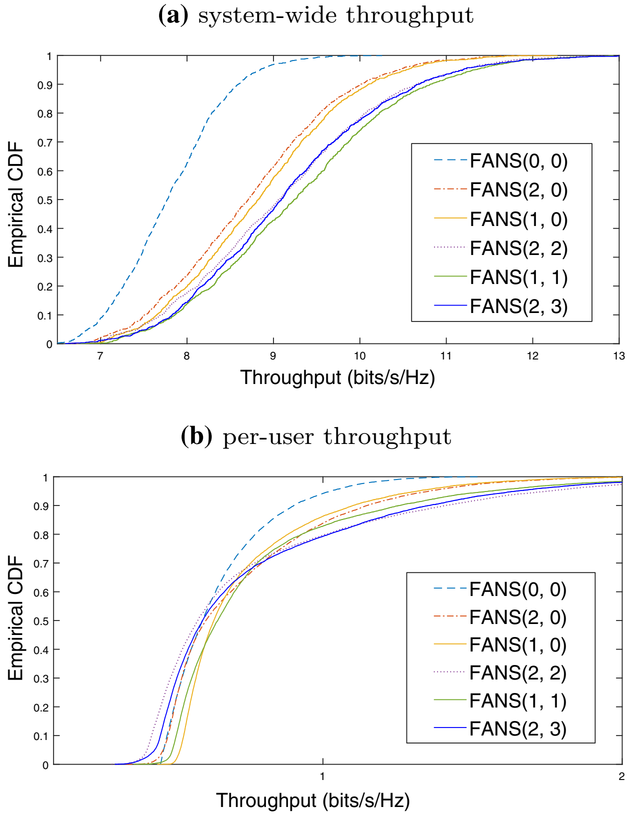
In this section, we provide various numerical examples to evaluate the performance of the proposed scheduler. Table 1 presents the parameters of the cellular system of interest used for obtaining the numerical results. The assume Rayleigh fading for all the users and the path loss and shadowing affects are also considered.

5.1 Impact of the scheduling parameter α

In the first numerical example, we study the impact of the scheduling parameter α in Algorithm 1 on the system-wide throughput $r(K)$ and the fairness index $JFI(K)$ over a span of K time slots. For this purpose, we fix $M = 12$, $R = 500$ m, transmit power $P = 40$ dBm and assume that all the user buckets are initially set to zero. Additionally, we assume that the twelve users are randomly placed in the cell and the three algorithms FANS(0,0), FANS(1,1), and FANS(2,2) are run for 10^6 time slots. The performance figures $r(K)$ and $JFI(K)$ are plotted as function of K in Fig. 3 from which we observe the following. For all the cases we had tried, Jain's fairness index $JFI(K)$ approaches to one as K grows, demonstrating long-term temporal fairness among users. However, the associated convergence rate is higher when α is large, and is lower when α is small. On the other hand, the overall system throughput is slightly lower in the steady-state for larger values of α for all the three versions of FANS. Also, we observe that for a fixed value of α, the system throughput figures attained by the NOMA-based FANS(1,1) and FANS(2,2) are significantly higher than that attained by OMA-based FANS(0,0). Furthermore, it can be seen that the Jain's fairness index converges faster when using FANS(1,1) and FANS(2,2) compared to FANS(0,0), favoring NOMA-based scheduling over the OMA-based scheme in terms of temporal fairness.

5.2 Impact of the proposed NOMA scheduler on system-wide and per-user throughputs

In this example, we investigate the impact of the proposed temporal fair NOMA scheduler on the system-wide and per-user throughputs. To this end, we consider 1000 system realizations and fix $M = 12$ users and algorithm parameter $\alpha = 0.01$. In each realization, we randomly place the twelve users in a cell of radius $R = 500$ m and run the six proposed scheduling algorithms for a duration of $K = 250000$ time-slots. The empirical cumulative distribution function (CDF) of the system-wide throughput and of the per-user throughput is plotted in Fig. 4. For the system-wide throughput results, we observe that the performance of FANS(2,0) that uses the reduced cardinality pair set is very close to that of FANS(1,0) that uses the entire pair set. Similarly, the system-wide performances of FANS(2,2) and FANS(2,3) that use the reduced cardinality pair and triple sets are very close to that of FANS(1,1) that uses the entire pair and triple sets. Additionally, we observe that FANS(2,3) slightly outperforms FANS(2,2) for instances with low system-wide throughputs.


(a) system-wide throughput

(b) per-user throughput

Fig. 4 Empirical CDF for: **a** system-wide throughput; **b** per-user throughput, over 1000 simulation instances each run for 250000 slots for the case $M = 12$, $P_T = 40$ dBm, $R = 500$ m

For the user-wide throughput results, we first examine the schemes which do not use triples of users, i.e. FANS(1,0) and FANS(2,0). In this case, weak (strong) users attain better (worse) throughput with FANS(1,0) in comparison with FANS(2,0) favoring FANS(1,0) over FANS(2,0) in terms of throughput fairness in addition to system-wide throughput. When triples of users are allowed, weak (strong) users attain better (worse) throughput with FANS(2,3) in comparison with FANS(2,2) favoring FANS(2,3) over FANS(2,2) in terms of throughput fairness. FANS(1,1) outperforms FANS(2,2) and FANS(2,3) in terms of throughput fairness but we note its relatively high computational complexity of $\mathcal{O}(M^3)$ compared to $\mathcal{O}(M)$ of the latter two scheduling algorithms. As a general observation, FANS(2,3) appears to be an appropriate choice trading off computational complexity, system-wide throughput, and throughput fairness.

Stemming from these observations, in the rest of the numerical examples, we will study only FANS(2,0), FANS(2,2),

Table 2 The steady-state system throughput r obtained with the four schedulers FANS(0,0), FANS(2,0), FANS(2,2), and FANS(2,3), for various values of the system parameters M, P, and R

		Transmit power P (dBm)					
		30			40		
		Cell radius R (meters)			Cell radius R (meters)		
M	Scheduler	100	500	1000	100	500	1000
18	FANS(0,0)	12.06	4.67	1.90	15.40	7.90	4.49
	FANS(2,0)	12.48	5.27	2.30	15.85	8.88	5.15
		(3.5%)	(12.8%)	(21.1%)	(2.9%)	(12.4%)	(14.7%)
	FANS(2,2)	12.59	5.63	2.63	16.00	9.29	5.59
		(4.4%)	(20.6%)	(38.4%)	(3.9%)	(17.6%)	(24.5%)
	FANS(2,3)	12.63	5.65	2.67	15.98	9.38	5.66
		(4.7%)	(21.0%)	(40.5%)	(3.8%)	(18.7%)	(26.1%)
36	FANS(0,0)	12.21	4.84	2.00	15.52	8.06	4.62
	FANS(2,0)	12.66	5.47	2.42	16.01	9.09	5.30
		(3.7%)	(13.0%)	(21.0%)	(3.2%)	(12.8%)	(14.7%)
	FANS(2,2)	12.80	5.83	2.77	16.19	9.51	5.73
		(4.8%)	(20.5%)	(38.5%)	(4.3%)	(18.0%)	(24.0%)
	FANS(2,3)	12.81	5.84	2.77	16.20	9.58	5.79
		(4.9%)	(20.6%)	(38.5%)	(4.4%)	(18.9%)	(25.3%)
72	FANS(0,0)	12.30	4.98	2.10	15.62	8.22	4.81
	FANS(2,0)	12.78	5.61	2.51	16.14	9.28	5.50
		(4.0%)	(12.7%)	(19.5%)	(3.3%)	(12.9%)	(14.3%)
	FANS(2,2)	12.92	5.96	2.87	16.33	9.70	5.95
		(5.0%)	(19.7%)	(36.7%)	(4.6%)	(18.0%)	(23.7%)
	FANS(2,3)	12.96	5.98	2.90	16.37	9.80	6.01
		(5.4%)	(20.0%)	(38.1%)	(4.8%)	(19.2%)	(25.0%)

The numbers inside the parentheses denote the percentage gain attained by the associated scheduler with respect to the OMA-based FANS(0,0) in terms of the system-wide throughput

and FANS(2,3) for NOMA-based user group scheduling in addition to the OMA-based FANS(0,0) due to their relative computational efficiency.

5.3 Impact of the number of users, transmit power, and cell radius

In this section, we examine the performance of the four schedulers, namely the OMA-based FANS(0,0), and NOMA-based FANS(2,0), FANS(2,2), and FANS(2,3), for three different values of $M \in \{18, 36, 72\}$, two different values of the transmit power $P \in \{30, 40\}$ dBm, and three different values of the cell radius $R \in \{100, 500, 1000\}$ m. As in the previous example, the results are obtained by averaging over 1000 system realizations each with a span of $K = 250000$ time slots. In each realization, M users are randomly placed in the cell and their bucket values are initialized to zero at the beginning. The estimate of the expected steady-state system throughput denoted by r is defined as

$$r = \frac{1}{L}\sum_{i=1}^{L} r^{(i)}(K),$$

where $r^{(i)}(K)$ is the system-wide throughput achieved at realization i. Table 2 lists the value of r for various combination of the number of users, cell radius, and BS transmit power. We have the following observations:

- The NOMA-based FANS(2,2) and FANS(2,3), which are making use of both pairs and triples, outperform the NOMA-based FANS(2,0) which is only making use of pairs. Additionally, all three NOMA-based schedulers outperform the OMA-based scheduler FANS(0,0). It is also observed that the relative performance improvement, i.e., with respect to FANS(0,0), of the NOMA-based schedulers increases with reduced BS transmit power and increased cell radius. The reason is when the transmit power is larger and/or the cell radius is smaller, the majority of the users will be strong and NOMA-based approaches would not be as effective.
- Out of the schedulers that can choose triples of users, FANS(2,3) outperforms FANS(2,2) slightly. In particular, we obtain up to 38.5% and 40.5% improvement, respectively, using FANS(2,2) and FANS(2,3), in system-wide throughput with respect to the OMA-based FANS(0,0) whereas with FANS(2,0), we obtain up to

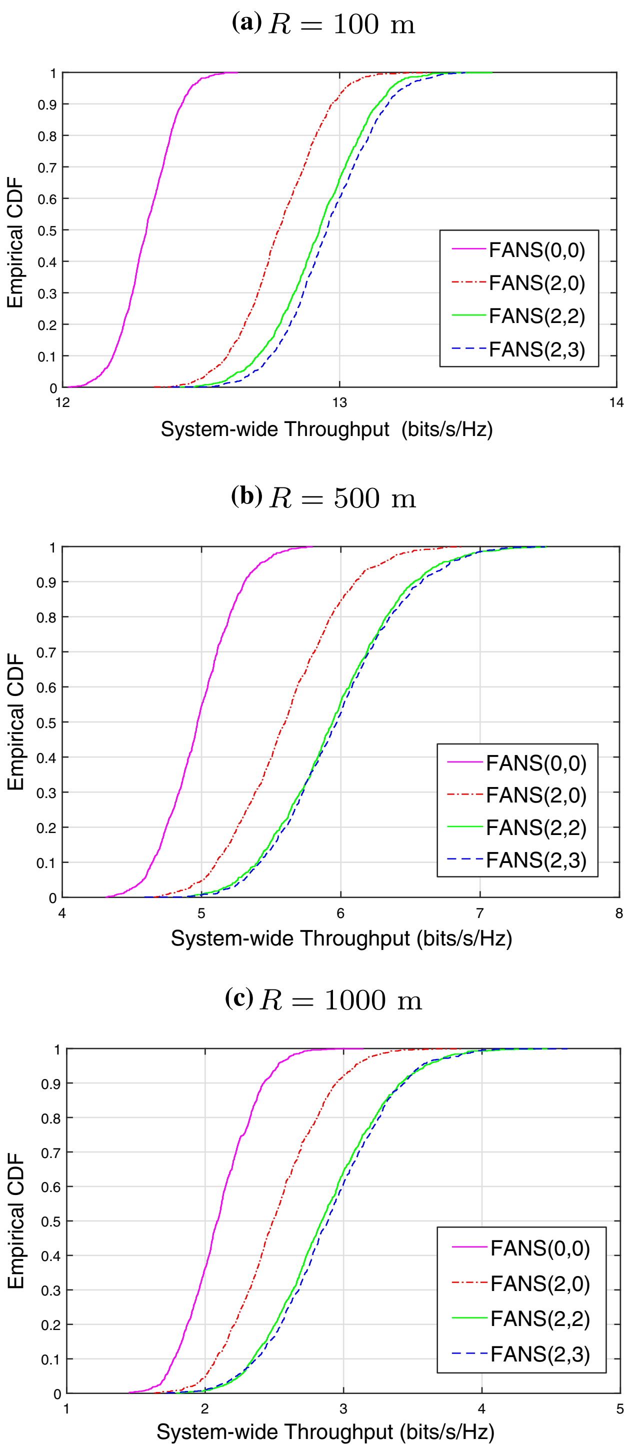


Fig. 5 Empirical CDF of the system-wide throughput obtained over 1000 simulation instances when $M = 72$, $P = 30$ dBm, and **a** $R = 100$ m, **b** $R = 500$ m, **c** $R = 1000$ m

21.1% improvement. These maximum gains are attained in NOMA-friendly regimes, i.e., relatively lower transmit power and larger cell radius, so that there is a rich mixture of strong and weak users that NOMA benefits from.

- When the number of users M increases, the system-wide throughput also increases slightly for all the schedulers due to higher multi-user diversity gains. However, the relative gains attained by using the NOMA-based schedulers with respect to the OMA-based FANS(0,0) do not noticeably change by M.

We also provide the empirical CDF of the system-wide throughput (over 1000 instances) in Fig. 5 when $M = 72$, $P = 30$ dBm, and for three different values of the cell radius R. While Table 2 confirms system throughput improvement on average (over various realizations) when using NOMA-based schedulers with respect to the OMA-based one, Fig. 5 reveals a consistent performance improvement for each realization. Furthermore, the performance gains of the NOMA-based schedulers with respect to the OMA-based scheduler tend to increase with increased radius as expected.

6 Conclusions

In this paper, we propose a number of opportunistic user group scheduling algorithms that dynamically select individual users, user pairs, or user triples, in a given cell, with the goal of maximizing the overall system throughput while ensuring temporal fairness among the contending users. Additionally, for computational efficiency purposes, we propose to dynamically construct reduced cardinality user pair and user triple sets as opposed to using the entire pair and triple sets, out of which a pair or a triple is to be selected. With the resulting NOMA-based scheduler FANS(2,3) that dynamically schedules pairs and triples out of reduced cardinality pair and triple sets, we have been able to obtain up to 40.5% improvement in system-wide throughput with respect to the OMA-based scheduler FANS(0,0). FANS(2,3) provides consistent performance improvement over FANS(0,0) but it is most effective in cases with larger cell radii and relatively smaller transmit powers. Our future work will consist of extension of our proposed algorithms to multi-carrier and MIMO-based NOMA systems and also taking into consideration per-source queue utilizations for scheduling purposes.

Funding None

Data availability To be made available upon request.

Declarations

Conflict of interest Authors declare that they have no conflict of interest.

newpage

References

1. Agiwal, M., Roy, A., & Saxena, N. (2016). Next generation 5G wireless networks: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, *18*(3), 1617–1655.
2. Al-Abbasi, Z.Q., & So, D.K. (2016). User-pairing based non-orthogonal multiple access (noma) system. In *2016 IEEE 83rd vehicular technology conference* (VTC Spring), IEEE, pp 1–5.
3. Aldababsa, M., Toka, M., Gokçeli, S., Kurt, G. K., & Kucur, O. (2018). A tutorial on nonorthogonal multiple access for 5G and beyond. *Wireless Communications and Mobile Computing*,. https://doi.org/10.1155/2018/9713450.
4. Ali, M. S., Tabassum, H., & Hossain, E. (2016). Dynamic user clustering and power allocation for uplink and downlink non-orthogonal multiple access (noma) systems. *IEEE Access*, *4*, 6325–6343.
5. Andrews, M., & Zhang, L. (2011). Scheduling algorithms for multicarrier wireless data systems. *IEEE/ACM Transactions on Networking*, *19*(2), 447–455.
6. Boccardi, F., Heath, R. W., Lozano, A., Marzetta, T. L., & Popovski, P. (2014). Five disruptive technology directions for 5G. *IEEE Communications Magazine*, *52*(2), 74–80.
7. Borgia, E., Bruno, R., Conti, M., Mascitti, D., & Passarella, A. (2016). Mobile edge clouds for information-centric IoT services. In *2016 IEEE Symposium on Computers and Communication (ISCC)*, IEEE, pp 422–428.
8. Choi, J. (2016). Power allocation for max-sum rate and max-min rate proportional fairness in NOMA. *IEEE Communications Letters*, *20*(10), 2055–2058.
9. Cisco. (2017). Cisco visual networking index: Global mobile data traffic forecast update, 2016–2021 white paper.
10. Cui, J., Ding, Z., & Fan, P. (2016). A novel power allocation scheme under outage constraints in NOMA systems. *IEEE Signal Processing Letters*, *23*(9), 1226–1230.
11. Dai, L., Wang, B., Ding, Z., Wang, Z., Chen, S., & Hanzo, L. (2018). A survey of non-orthogonal multiple access for 5G. *IEEE Communications Surveys Tutorials*, *20*, 2294.
12. Ding, Z., Lei, X., Karagiannidis, G.K., Schober, R., Yuan, J., & Bhargava, V. (2017). A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends. arXiv preprint arXiv:170605347.
13. Fang, F., Zhang, H., Cheng, J., Roy, S., & Leung, V. C. (2017). Joint user scheduling and power allocation optimization for energy-efficient NOMA systems with imperfect CSI. *IEEE Journal on Selected Areas in Communications*, *35*(12), 2874–2885.
14. Islam, S. M. R., Avazov, N., Dobre, O. A., & Kwak, K. (2017). Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges. *IEEE Communications Surveys Tutorials*, *19*(2), 721–742.
15. Jain, R. (1991). *The art of computer systems performance analysis-techniques for experimental design, measurement, simulation, modeling*. Hoboken: Wiley.
16. Jalali, A., Padovani, R., & Pankaj, R. (2000). Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system. In *VTC2000-Spring. 2000 IEEE 51st Vehicular Technology Conference Proceedings*, *3*, 1854–1858.
17. Lei, L., Yuan, D., & Värbrand, P. (2016). On power minimization for non-orthogonal multiple access (NOMA). *IEEE Communications Letters*, *20*(12), 2458–2461.
18. Li, A., Benjebbour, A., & Harada, A. (2014). Performance evaluation of non-orthogonal multiple access combined with opportunistic beamforming. In *Vehicular Technology Conference (VTC Spring)*, 2014 IEEE 79th, IEEE, pp 1–5.
19. Liu, F., & Petrova, M. (2017). Proportional fair scheduling for downlink single-carrier NOMA systems. In *GLOBECOM 2017–2017 IEEE Global Communications Conference*, pp 1–7.
20. Liu, X., Chong, E. K. P., & Shroff, N. B. (2001). Opportunistic transmission scheduling with resource-sharing constraints in wireless networks. *IEEE Journal on Selected Areas in Communications*, *19*(10), 2053–2064.
21. Liu, X., Chong, E. K., & Shroff, N. B. (2003). A framework for opportunistic scheduling in wireless networks. *Computer Networks*, *41*(4), 451–474.
22. Morgado, A., Huq, K. M. S., Mumtaz, S., & Rodriguez, J. (2018). A survey of 5G technologies: Regulatory, standardization and industrial perspectives. *Digital Communications and Networks*, *4*(2), 87–97.
23. Oviedo, J. A., & Sadjadpour, H. R. (2017). A fair power allocation approach to NOMA in multiuser SISO systems. *IEEE Transactions on Vehicular Technology*, *66*(9), 7974–7985.
24. Parida, P., & Das, SS. (2014). Power allocation in OFDM based NOMA systems: A DC programming approach. In *2014 IEEE Globecom Workshops (GC Wkshps)*, IEEE, pp 1026–1031.
25. Saito, Y., Kishiyama, Y., Benjebbour, A., Nakamura, T., Li, A., & Higuchi, K. (2013). Non-orthogonal multiple access (NOMA) for cellular future radio access. In *Vehicular Technology Conference (VTC Spring)*, 2013 IEEE 77th, IEEE, pp 1–5.
26. Shahab, M. B., Irfan, M., Kader, M. F., & Young Shin, S. (2016a). User pairing schemes for capacity maximization in non-orthogonal multiple access systems. *Wireless Communications and Mobile Computing*, *16*(17), 2884–2894.
27. Shahab, M. B., Kader, M. F., & Shin, S. Y. (2016b). A virtual user pairing scheme to optimally utilize the spectrum of unpaired users in non-orthogonal multiple access. *IEEE Signal Processing Letters*, *23*(12), 1766–1770.
28. Shahsavari, S. (2019). Access, resource allocation, and performance analysis in next generation wireless networks. PhD thesis, New York University Tandon School of Engineering.
29. Shahsavari, S., & Akar, N. (2015). A two-level temporal fair scheduler for multi-cell wireless networks. *IEEE Wireless Communications Letters*, *4*(3), 269–272.
30. Shahsavari, S., Akar, N., & Hossein, B. (2018). Joint cell muting and user scheduling in multicell networks with temporal fairness. *Mobile Communications and Wireless Networking*,. https://doi.org/10.1155/2018/4846291.
31. Shahsavari, S., Shirani, F., & Erkip, E. (2018). Opportunistic temporal fair scheduling for non-orthogonal multiple access. In *2018 56th Annual Allerton conference on communication, control, and computing (Allerton)*, pp 391–398.
32. Shahsavari, S., Shirani, F., Amir Khojastepour, MA., & Erkip, E. (2019a). Opportunistic temporal fair mode selection and user scheduling for full-duplex systems. In *2019 IEEE 30th international symposium on personal, indoor and mobile radio communications (PIMRC Workshops)*, pp 1–7.
33. Shahsavari, S., Shirani, F., & Erkip, E. (2019b). A general framework for temporal fair user scheduling in NOMA systems. *IEEE Journal of Selected Topics in Signal Processing*, *13*(3), 408–422.
34. Shahsavari, S., Shirani, F., & Erkip, E. (2019c). On the fundamental limits of multi-user scheduling under short-term fairness constraints. In 2019 *IEEE international symposium on information theory (ISIT)*, pp 2534–2538.
35. Tassiulas, L., & Ephremides, A. (1992). Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *IEEE Transactions on Automatic Control*, *37*(12), 1936–1948.
36. Tse, D., & Viswanath, P. (2005). *Fundamentals of Wireless Communication*. New York, NY, USA: Cambridge University Press.
37. Xin Liu, Chong, EKP., & Shroff, NB. (2001). Transmission scheduling for efficient wireless utilization. In *Proceedings IEEE*

INFOCOM 2001. Conference on Computer Communications. Twentieth Annual Joint Conference of the IEEE Computer and Communications Society, vol 2, pp 776–785.

38. Yang, L., Chen, J., Ni, Q., Shi, J., & Xue, X. (2017). NOMA-enabled cooperative unicast-multicast: Design and outage analysis. *IEEE Transactions on Wireless Communications*, *16*(12), 7870–7889.
39. Yang, Z., Xu, W., Pan, C., Pan, Y., & Chen, M. (2017). On the optimality of power allocation for NOMA downlinks with individual QoS constraints. *IEEE Communications Letters*, *21*(7), 1649–1652.
40. Zeng, M., Yadav, A., Dobre, O. A., Tsiropoulos, G. I., & Poor, H. V. (2017). Capacity comparison between MIMO-NOMA and MIMO-OMA with multiple users in a cluster. *IEEE Journal on Selected Areas in Communications*, *35*(10), 2413–2424.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Eray Erturk received his B.S. degree in Electrical and Electronics Engineering from Bilkent University, Ankara, Turkey, in 2020. He is currently a Ph.D. student in the Ming Hsieh Department of Electrical and Computer Engineering at the University of Southern California. He received a Viterbi Fellowship from USC in 2020 and currently, he is working on brain studies through modelling, decoding and control of neural dynamics in the Neural Systems Engineering & Information Processing (NSEIP) Lab.

Ozlem Yildiz received her B.S. degree in Electrical and Electronics Engineering from Bilkent University, Ankara, Turkey, in 2020. She is currently a Ph.D. student in the Department of Electrical and Computer Engineering at NYU Tandon School of Engineering. Her research interests are on wireless communications.

Shahram Shahsavari received the B.S. degree in electrical engineering from Amirkabir University of Technology (Tehran Polytechnic), Iran, in 2013, and the M.S. degree in electrical engineering from Sharif University of Technology, Iran, in 2015. He received the Ph.D. degree in electrical engineering from the New York University Tandon School of Engineering, NY, USA, in 2019. During his Ph.D, he was also with NYU WIRELESS Center at New York University conducting research on next generation wireless networks, and was the recipient of the NYU Ernst Weber fellowship in 2015. He joined University of Waterloo in 2019 as a postdoctoral research fellow to conduct research on various technologies for 5G and beyond. His research interests include wireless communications, modern cellular systems, radio resource management, and network optimization.

Nail Akar received his B.S. degree from Middle East Technical University, Turkey, in 1987 and M.S. and Ph.D. degrees from Bilkent University, Ankara, Turkey, in 1989 and 1994, respectively, all in electrical and electronics engineering. From 1994 to 1996, he was a visiting scholar and a visiting assistant professor in the Computer Science Telecommunications program at the University of Missouri - Kansas City, USA. He joined the Technology Planning and Integration group at Long Distance Division, Sprint, Overland Park, Kansas, in 1996, where he held a senior member of technical staff position from 1999 to 2000. Since 2000, he has been with Bilkent University currently as a Professor of the Electrical and Electronics Engineering Department and as the Associate Dean of the Engineering Faculty. He visited the School of Computing, University of Missouri - Kansas City, as a Fulbright scholar in 2010 for a period of six months. His research interests are on performance modeling of computer and communication systems and networks, wireless networks, Internet of Things, queueing theory, and optimization.