Design and Optimization of Scheduling and Non-orthogonal Multiple Access Algorithms with Imperfect Channel State Information

Jianhua He, *Senior Member, IEEE*, Zuoyin Tang, Zuowen Tang, Hsiao-Hwa Chen, *Fellow, IEEE*, and Cong Ling

***Abstract*—Non-orthogonal multiple access (NOMA) is a promising candidate technology for 5G cellular systems. In this paper, design and optimization of scheduling and NOMA algorithms is investigated. The impact of power allocation for NOMA systems with round-robin scheduling is analyzed. A statistic model is developed for network performance analysis of joint scheduling of spectrum resource and power for NOMA algorithms. Proportional fairness (PF) scheduling for NOMA is proposed with a two-step approach, with objective of achieving high throughput and user fairness with low computational complexity. In the first step, an optimal power allocation strategy is developed with an objective of maximizing weighted sum rate. In the second step, three fast and scalable scheduling and user pairing algorithms with QoS guarantee are proposed, in which only a few user pairs are checked for NOMA multiplex. The algorithms are extended to the cases with imperfect channel state estimation and more than two users being multiplexed over one resource block. Numerical results show that the proposed algorithms are faster and more scalable than the existing algorithms, and maintain a higher throughput gain than orthogonal multiple access.**

***Index Terms*—Non-orthogonal multiple access; Scheduling; Cellular network; 5G; Power allocation; Cross layer design**

I. INTRODUCTION

Non-orthogonal multiple access (NOMA) is a promising technology currently under consideration for 5G systems [1]–[5]. In an orthogonal multiple access (OMA) system, such as orthogonal frequency division multiple access (OFDMA), frequency-time resource is allocated exclusively to at most one user equipment (UE) in the same cell. NOMA systems allow simultaneous allocation of the same frequency resource to multiple UEs in the same cell, offering a superior spectral efficiency and massive connectivity [2]–[4].

Non-orthogonal resource allocation and signal reception can be achieved in power and code domains [5]–[8]. This paper focuses on the power domain NOMA, where multiple UEs can be multiplexed in the power domain with superposition coding [10], [11]. Intra-cell interference at a receiver is cancelled with successive interference cancellation (SIC) [2], [10]–[12]. With the help of NOMA, more than 20% throughput gain was reported in the literature [2]–[4]. An illustration of NOMA network with superposition coding and SIC is shown in Fig.1.

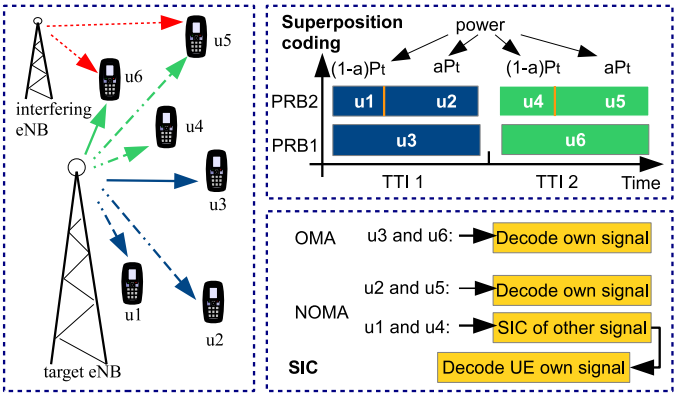


Fig. 1. Illustration of network operation with scheduling and NOMA. Six UEs are served by a target eNB over two transmit time intervals (TTI) subject to neighbor eNB interference. u_1, u_2, and u_3 are scheduled in TTI 1. u_1 and u_2 are multiplexed with superposition coding over PRB 1. u_1 decodes its own signal after cancelling u_2 signal from received superimposed signal, and u_2 decodes its own signal directly.

A. Related Works

The promising performance of NOMA stimulated a lot of research efforts. Detailed literature surveys on the recent NOMA research can be found in [7]–[9]. These research works can be classified to two main categories, i.e., theoretic modeling and simulation approaches.

The early theoretic model based research works were focused on the evaluation of NOMA performance gain over OMA [2]–[4], [13]. Later on, there are research works reported on the design of NOMA with limited feedback [14], and integration with complementary wireless technologies, e.g., multiple input multiple output (MIMO), beamforming, relaying, device to device communication [15]–[19], and vehicular networks [20]. It is noted that, while theoretic models provide analytical tools for NOMA algorithmic design, system models and research methodology were somehow over-simplified in these works. For example, a large number of analytical works assumed system models with two users or grouped users [13], [15], [31], [32]. Inter-cell interference was ignored in the above works for the sake of model tractability. Network

Jianhua He (email: j.he7@aston.ac.uk), Zuoyin Tang (email: z.tang1@aston.ac.uk), and Zuowen Tang (email: tangz3@aston.ac.uk) are with the School of Engineering and Applied Science, Aston University, UK. Hsiao-Hwa Chen (email: hshwchen@mail.ncku.edu.tw) is with the Department of Engineering Science, National Cheng Kung University, Taiwan. Cong Ling (email: c.ling@imperial.ac.uk) is with the Department of Electrical and Electronic Engineering, Imperial College London, UK.

This manuscript was submitted on February 25, 2018 and revised on June 13, 2018.

level scheduling and user grouping and user fairness were not considered. These simplified treatments may limit the scope and validity of the NOMA algorithm design and the insights obtained from the above research works.

For the simulation based NOMA algorithmic design and evaluation research works, their focus is on resource allocation, which is a core research problem in NOMA networks. In addition to traditional research problems of allocating spectrum and power resources, there is an extra dimension of user pairing for NOMA resource allocation. In the early NOMA research works [2]–[4], some heuristic strategies were proposed for power allocation, and an exhaustive search (ES) approach was used for scheduling and user pairing. The computational complexity of ES approach can be prohibitively high for both simulations and practical NOMA applications. Parida *et al.* [21] proposed a greedy user selection algorithm and a difference of convex (DC) programming based power allocation algorithm for NOMA resource allocation. Proportional fairness (PF) scheduling algorithms were used in [2], [21]. In [22], user selection and power allocation for NOMA beamforming systems was investigated, but scheduling over time was not considered and ES approach was used for user selection.

Fang *et al.* [23] proposed joint subchannel and transmit power allocation to maximize NOMA network energy efficiency. Matching theory was applied to subchannel allocation, and DC programming was applied to power allocation within and across subchannels. The joint subchannel and power allocation problem with imperfect channel state information (CSI) was investigated in [24]. Sun *et al.* [25] proposed a successive convex approximation based joint subcarrier and power allocation algorithm for full-duplex NOMA systems. Wei *et al.* [26] applied DC programming to design an iterative resource allocation algorithm to maximize optimal energy efficiency, in which imperfect CSI was taken into account in the algorithm design. Zhu *et al.* [27] investigated matching based channel assignment and optimal power allocation for NOMA systems. Performance optimization criteria including maximin fairness and weighted sum rate maximization were considered with QoS constraints.

B. Motivation and Contributions

Resource allocation and user scheduling are two core components of 5G cellular systems, which are illustrated in Fig. 1. User scheduling and NOMA will be expected to coexist in 5G cellular systems if NOMA is adopted. Additional user pairing and power allocation (UPPA) for NOMA makes the existing scheduling and resource allocation problems more complicated.

While some interesting research works on NOMA resource allocation and scheduling have been reported, the research is still in its early stage and there are many open research issues. For example, the latest simulation based resource allocation works considered only a single cell scenario [22]–[27]. Continuous scheduling (e.g., using PF scheduling algorithm) and user fairness were not considered in these works. The computational complexity and algorithm running time were not evaluated. In [23], [24], [26], the main performance metric of interest is energy efficiency, and the network throughput performance was not considered. For matching theory based user selection algorithms, which were used in many works such as [23], [27], the two-to-one matching sets an over strict limitation on the applicability of algorithms as the number of users is required to be twice the number of the resource channels.

The full potentials of NOMA can only be realized with properly designed resource allocation and scheduling algorithms. The joint scheduling with NOMA UPPA algorithms need to be effective and fast, and be evaluated in more realistic network scenarios. While suboptimal algorithms were proposed to reduce computational complexity in [22]–[27], the effectiveness and speed of the algorithms still require much more investigations. Practical scheduling with PF algorithm was applied in [2], [21], but the computational complexity of their power allocation and scheduling algorithms is still too high.

In view of the aforementioned research gaps, in this paper we aim to develop fast and effective joint scheduling with NOMA algorithms on top of our preliminary works [28] [29]. Unlike the aforementioned research works, this paper considers practical network settings, such as inter-cell interference, user QoS constraints, imperfect CSI estimation, and practical scheduling algorithms. Several power allocation strategies and user scheduling algorithms are proposed and evaluated, with their design objectives of minimizing algorithm computational complexity and maintaining a good network performance in terms of network throughput and user fairness. As round-robin (RR) scheduling and proportional fairness (PF) scheduling are two widely used scheduling algorithms, they are chosen in this study on joint design of scheduling and NOMA algorithms.

II. SYSTEM MODEL

Consider a cellular network with N_{site} sites, each equipped with one eNodeB (eNB). The eNBs are labelled from 1 to N_{site}. eNB 1 is located at the network center. Each eNB has three sectors. Each sector represents a cell. The jth sector of the ith site is denoted by $\mathcal{A}_{i,j}$, where $i \in [1, N_{\text{site}}]$ and $j \in [1, 3]$. A clover-leaf network layout is used, which shows a better performance than a hexagonal network layout [35]. UEs are assumed to be randomly and uniformly distributed in a network service area. Let Ω_{ue} denote a set of UEs. A full buffer traffic model is assumed. Due to the symmetry of the sector structure and the full load traffic assumption, it is expected that all sectors have very similar performances. Therefore, the analysis of UEs in a representative sector (i.e., sector $\mathcal{A}_{1,1}$ in this paper) is sufficient for system-level performance evaluation. Assume that there are N_{rb} physical resource blocks (PRBs). PRB represents basic time-frequency resource unit for data transmission in LTE networks.

Table I lists the main notations used in this paper, where superscripts "o", "m" and "n" in variables are designated OMA, NOMA multiplexing, and NOMA, respectively.

TABLE I
NOTATIONS AND THEIR DEFINITIONS.

Notation	Definition	Notation	Definition
N_{site}	Number of sites	$\mathcal{A}_{i,j}$	The jth sector of the ith site
P_{t}	eNB transmit power over one PRB	$\mathcal{P}_{i,j,u,r}$	Power received by UE u from $\mathcal{A}_{i,j}$ over PRB r
$P_{\text{r,i,j,u}}$	Mean received power by u from $\mathcal{A}_{i,j}$	$\psi_{i,u}$	Shadow fading between eNBs i and u
σ_{w}	Log-normal shadowing standard deviation	ρ	Inter-site shadow fading correlation
N_{m}	Maximal number of UEs sharing a PRB	α	NOMA power allocation coefficient (PAC)
$\gamma^{\text{o}}_{u,r}$	OMA SIR of UE u	$\gamma^{\text{m}}_{u_1,r}(u_1,u_2,\alpha)$	SIR of u_1 when multiplexed with u_2
C^{o}_{u}	OMA spectral efficiency (SE) of u	$C^{\text{m}}_{u_1,r}(u_1,u_2,\alpha)$	SE of u_1 when multiplexed with u_2
$\overline{C^{\text{n}}_{\text{net}}}(\alpha)$	Mean network SE with α	$\eta_{\text{site}}(\alpha)$	Mean site throughput with α
$I_{i,u}$	Aggregate interference from eNBs i to u	$I_{\text{a},u}$	Aggregate interference from all eNBs to u
$\varphi_{u,r,t}(\alpha)$	Priority coefficient (PC) of UE u over PRB r	$\varphi_{u_1,u_2,r,t}(\alpha)$	Sum PC of multiplexed u_1 and u_2 over r

A. Channel Model and Antenna Radiation Pattern

Let $\mathcal{P}_{i,j,u,r}$ be signal power received by a UE u from sector $\mathcal{A}_{i,j}$ over PRB r, which is computed by

$$\mathcal{P}_{i,j,u,r} = P_{\text{t}} G_{\text{PL}}(i,u) G_{\text{A}}(i,j,u) \psi_{i,u} \phi_{i,j,u,r}, \tag{1}$$

where P_{t} denotes eNB transmission power over one PRB, $G_{\text{PL}}(i,u)$ is a path gain between eNB i and UE u, $G_{\text{A}}(i,j,u)$ denotes antenna gain between sector $\mathcal{A}_{i,j}$ and u, $\psi_{i,u}$ represents shadow fading between eNB i and u, and $\phi_{i,j,u,r}$ denotes small scale fast fading between $\mathcal{A}_{i,j}$ and u over PRB r. For ease of notation, let $P_{\text{r,i,j,u}} = P_{\text{t}} G_{\text{PL}}(i,u) G_{\text{A}}(i,j,u)$ be the received power at UE u from sector $\mathcal{A}_{i,j}$ without fading.

The path gain (loss) $G_{\text{PL}}(d)$ models the propagation loss between eNB i and UE u. The model specified in [34] for outdoor line-of-sight communications is used, or

$$G_{\text{PL}}(i,u) = -103.4 - 24.2 \log_{10}(d_{i,u}) \text{ (dB)}, \tag{2}$$

where $d_{i,u}$ is the distance in kilometers between eNB i and UE u.

Shadow fading $\psi_{i,u}$ between eNB i and UE u is assumed to follow a log-normal distribution with zero mean and standard deviation of σ_{w} [34]. Moreover, the shadow fading within sectors of a site is assumed to be fully correlated, while the inter-site shadow fading correlation is denoted by ρ. The antenna gain $G_{\text{A}}(i,j,u)$ models the gain of an antenna in the direction between sector $\mathcal{A}_{i,j}$ and UE u. The same antenna model and parameter settings for the model used in [35] are applied in this paper, which are not repeated here.

B. SIR for UEs with OMA and NOMA

In this subsection, let us consider a signal to interference (SIR) model for a UE serviced by OMA and a pair of UEs serviced by NOMA, which provides a basis for statistical network performance analysis of NOMA systems in Section III and the design of NOMA UPPA algorithms in Section IV.

In the OMA systems, a PRB is allocated to at most one UE (say u) in one sector. UE u receives no intra-cell interference. Let $\gamma^{\text{o}}_{u,r}$ denote the SIR of UE u of sector $\mathcal{A}_{1,1}$ over PRB r with OMA (superscript o denotes OMA), which is computed as

$$\gamma^{\text{o}}_{u,r} = \frac{\mathcal{P}_{1,1,u,r}}{\sum_{j=2}^{3} \mathcal{P}_{1,j,u,r} + \sum_{i=2}^{N_{\text{site}}} \sum_{j=1}^{3} \mathcal{P}_{i,j,u,r}}. \tag{3}$$

As the downlink communication is assumed to be interference limited, noise power is negligible and not considered. In a NOMA system, according to the channel conditions of UEs, a PRB r may be allocated to more than one UE, but it is not mandatory. If a PRB is allocated to only one UE, UE SIR can be computed by (3). Initially, the maximal number of UEs that can be multiplexed over a PRB (denoted by N_{m}) is limited to two. The limitation is relaxed later in the design of UPPA algorithms.

If PRB r is allocated to two multiplexed UEs (say u_1 and u_2), according to NOMA principle, at the eNB side the desired signals targeting at u_1 and u_2 are superimposed over PRB r. Transmission powers $(1-\alpha)P_{\text{t}}$ and αP_{t} are allocated to the two UEs with a larger and a smaller OMA SIR, respectively. α is called power allocation coefficient (PAC). It is noted that a necessary condition on α is $\alpha > 0.5$; otherwise SIC at the UE with a lower OMA SIR is thought to fail, under an SIC assumption that a received signal cannot be successfully decoded and cancelled with SINR¡1. At the receiver side, UE with a poorer channel condition decodes its signal directly without SIC, by which the signal for the UE with a larger OMA SIR is treated as intra-cell UE interference. Let $\gamma^{\text{m}}_{u_1,r}(u_1,u_2,\alpha)$ and $\gamma^{\text{m}}_{u_2,r}(u_1,u_2,\alpha)$ denote respectively the SIRs of two multiplexed UEs u_1 and u_2, over PRB r with PAC α, under the condition $\gamma^{\text{o}}_{u_1,r} > \gamma^{\text{o}}_{u_2,r}$. Setting α is investigated in the subsequent sections. The superscript m designates intra-cell UE multiplexing.

The SIR $\gamma^{\text{m}}_{u_2,r}(u_1,u_2,\alpha)$ of UE u_2 is then computed by

$$\begin{aligned}
&\gamma^{\text{m}}_{u_2,r}(u_1,u_2,\alpha) \\
&= \frac{\alpha \mathcal{P}_{1,1,u_2,r}}{\sum_{j=2}^{3} \mathcal{P}_{1,j,u_2,r} + \sum_{i=2}^{N_{\text{site}}} \sum_{j=1}^{3} \mathcal{P}_{i,j,u_2,r} + (1-\alpha)\mathcal{P}_{1,1,u_2,r}} \\
&= \frac{1+\gamma^{\text{o}}_{u_2,r}}{1+(1-\alpha)\gamma^{\text{o}}_{u_2,r}}.
\end{aligned} \tag{4}$$

At UE u_1, intra-cell interference from u_2 is decoded and cancelled before u_1 desired signal is decoded. The SIR of UE u_1 is computed by

$$\begin{aligned}
\gamma^{\text{m}}_{u_1,r}(u_1,u_2,\alpha) &= \frac{(1-\alpha)\mathcal{P}_{1,1,u_1,r}}{\sum_{j=2}^{3} \mathcal{P}_{1,j,u_1,r} + \sum_{i=2}^{N_{\text{site}}} \sum_{j=1}^{3} \mathcal{P}_{i,j,u_1,r}} \\
&= (1-\alpha)\gamma^{\text{o}}_{u_1,r}.
\end{aligned} \tag{5}$$

It is noted that if $\gamma^{\mathrm{o}}_{u_1,r} \leq \gamma^{\mathrm{o}}_{u_2,r}$, the SIRs of the two multiplexed UEs, u_1 and u_2, can be computed as $\gamma^{\mathrm{m}}_{u_1,r}(u_2,u_1,\alpha)$ and $\gamma^{\mathrm{m}}_{u_2,r}(u_2,u_1,\alpha)$, respectively.

III. CROSS-LAYER DESIGN FOR NOMA WITH RR SCHEDULER

In this section, cross-layer design of an NOMA system with RR scheduler is investigated. RR scheduling is simple and easy to implement. In a traditional OMA system with RR scheduling, frequency-time resources are assigned to each UE equally in a circular order. In the considered NOMA systems with RR scheduling, their operations are loosely coupled and performed on a full set of UE pairs. In the full set of UE pairs, there is one and only one pair for each UE and every UE (including the UE itself) in the network. Instead of scheduling individual UE in an OMA system, the RR scheduler in a NOMA system schedules individual UE pair in the full UE pair set in a circular order. Each UE pair has an equal share of the PRBs. For each scheduled UE pair, NOMA multiplexing is not mandatory. If the channel condition of UEs is not desirable for NOMA multiplexing, two UEs are served by OMA, and each receives an half share of the frequency resource allocated to the scheduled UE pair. The sum rates of the paired UEs with NOMA and OMA are used in the decision making.

Due to the simplicity of RR scheduling and the loose coupling of NOMA and RR scheduling, the joint scheduling for NOMA design problem is reduced to the selection of NOMA multiplexing or OMA for a scheduled UE pair and power allocation for a multiplexed UE. Next, an analysis on the impact of power allocation in a NOMA system with RR scheduling is presented in Section III-A. The SIR distribution for a given pair of multiplexed UEs with fixed locations is derived in Section III-B. Analytical models for spectral efficiency (SE) of a scheduled UE pair, network throughput, and fairness are developed in Sections III-C and III-D, respectively.

A. Impact of Power Allocation

Power allocation plays an important role in NOMA systems. In a NOMA system with RR scheduling, there is no easy method to determine PAC α for NOMA. RR scheduling provides excellent UE fairness on network resource utilization. With the introduction of NOMA, resource utilization fairness will be significantly affected, i.e., a smaller α can increase the throughput of a network and the UE with a better channel quality, but gives worse UE fairness.

Let us consider two generic UEs, u_1 and u_2, to be multiplexed by NOMA. The UEs have SIRs $\gamma^{\mathrm{o}}_{u_1,r}$ and $\gamma^{\mathrm{o}}_{u_2,r}$ over PRB r. Without loss of generality, assume $\gamma^{\mathrm{o}}_{u_1,r} > \gamma^{\mathrm{o}}_{u_2,r}$. With Shannon capacity formula, the sum SE of u_1 and u_2 with NOMA multiplexing over PRB r is computed as

$$\begin{aligned}
&\log_2\big[1+\gamma^{\mathrm{m}}_{u_1,r}(u_1,u_2,\alpha)\big]+\log_2\big[1+\gamma^{\mathrm{m}}_{u_2,r}(u_1,u_2,\alpha)\big]\\
&=\log_2\big[1+(1-\alpha)\gamma^{\mathrm{o}}_{u_1,r}\big]+\log_2\Big[1+\frac{1+\gamma^{\mathrm{o}}_{u_2,r}}{1+(1-\alpha)\gamma^{\mathrm{o}}_{u_2,r}}\Big]\\
&=\log_2\Big[(1+\gamma^{\mathrm{o}}_{u_2,r})\big(1+\frac{\gamma^{\mathrm{o}}_{u_1,r}-\gamma^{\mathrm{o}}_{u_2,r}}{\frac{1}{1-\alpha}+\gamma^{\mathrm{o}}_{u_2,r}}\big)\Big]. \qquad (6)
\end{aligned}$$

It can be observed from (6) that the sum SE with NOMA multiplexing decreases monotonically with α. A higher power should be allocated to the UE with a better channel quality (i.e., α should be very close to 0.5) to maximize network throughput, but it is not fair in terms of resource utilization. Therefore, for a NOMA system with RR scheduling, PAC α is a system design parameter to be considered with both network throughput and fairness. An analytical model is proposed next to compute network throughput and fairness to support the control of α.

B. SIR PDF of Two Multiplexed UEs with Fixed Locations

Given a specific pair of UEs, u_1 and u_2, with their fixed locations, let $C^{m}_{u_1,u_2,\alpha}$ denote the sum SE of multiplexed UEs, u_1 and u_2, with α. As SIR probability density function and the mean SE of UEs in a NOMA system with RR scheduling are computed over all channel fading instantiations and PRBs, subscript r is not included in the new variables introduced in the remaining of Section III. Next, we derive a formula to compute $C^{m}_{u_1,u_2,\alpha}$ of the multiplexed UEs for the case of $\gamma^{\mathrm{o}}_{u_1,r} > \gamma^{\mathrm{o}}_{u_2,r}$. The sum SE in the case of $\gamma^{\mathrm{o}}_{u_1,r} \leq \gamma^{\mathrm{o}}_{u_2,r}$ can be computed similarly.

Let $I_{i,u}$ be the aggregate interference generated from all sectors of the ith eNB to a UE u, i.e.,

$$I_{i,u}=\begin{cases}\sum_{j=2}^{3} P_{\mathrm{r},1,j,u}\psi_{1,u}, & i=1,\\ \sum_{j=1}^{3} P_{\mathrm{r},i,j,u}\psi_{i,u}, & i=2,...,N_{\mathrm{site}}.\end{cases} \qquad (7)$$

Note that $I_{i,u}$ is not a log-normal random variable but is approximated as a log-normal variable with the method proposed in [33]. Let $\mu_{I_{i,u}}$ and $\sigma_{I_{i,u}}$ denote the mean and the standard deviation of a normal distribution associated with $I_{i,u}$, respectively, which can be calculated by

$$\mu_{I_{i,u}}=\begin{cases}\ln\Big(\sum_{j=2}^{N_{\mathrm{a}}} P_{\mathrm{r},i,j,u}\Big), & i=1,\\ \ln\Big(\sum_{j=1}^{N_{\mathrm{a}}} P_{\mathrm{r},i,j,u}\Big), & i=2,...,N_{\mathrm{site}},\end{cases} \qquad (8)$$

and

$$\sigma_{I_{i,u}}=\sigma_{\mathrm{w}}. \qquad (9)$$

According to (5), the SIR of UE u_1, $\gamma^{\mathrm{m}}_{u_1,r}(u_1,u_2,\alpha)$, can be expressed by

$$\begin{aligned}
\gamma^{\mathrm{m}}_{u_1,r}(u_1,u_2,\alpha)&=\frac{(1-\alpha)P_{\mathrm{r},1,1,u_1}\psi_{1,u_1}}{\sum_{j=2}^{3} P_{\mathrm{r},1,j,u_1}\psi_{1,u_1}+\sum_{i=2}^{N_{\mathrm{site}}}\sum_{j=1}^{3} P_{\mathrm{r},i,j,u_1}\psi_{i,u_1}}\\
&=\frac{(1-\alpha)P_{\mathrm{r},1,1,u_1}\psi_{1,u_1}}{\sum_{i=1}^{N_{\mathrm{site}}} I_{i,u_1}}. \qquad (10)
\end{aligned}$$

Note that fast fading is not included in the above formula as its impact is negligible in the analysis with shadow fading [35].

According to (4), the SIR of UE u_2, $\gamma^{\mathrm{m}}_{u_2,r}(u_1,u_2,\alpha)$, can be expressed similarly as

$$\gamma^{\mathrm{m}}_{u_2,r}(u_1,u_2,\alpha)=\frac{\alpha P_{\mathrm{r},1,1,u_2}\psi_{1,u_2}}{\sum_{i=1}^{N_{\mathrm{site}}} I_{i,u_2}+(1-\alpha)P_{\mathrm{r},1,1,u_2}\psi_{1,u_2}}. \tag{11}$$

As intra-site fading is assumed to be fully correlated and inter-site fading is partially correlated, intra-site and inter-site interferences are treated separately. Two intra-site interference related variables, $Y^{\mathrm{m}}_{u_1}$ and $Y^{\mathrm{m}}_{u_2}$, are introduced for multiplexed UEs, u_1 and u_1, which are computed by

$$Y^{\mathrm{m}}_{u_1}=\frac{\sum_{j=2}^{3} P_{\mathrm{r},1,j,u_1}}{(1-\alpha)P_{\mathrm{r},1,1,u}}, \tag{12}$$

$$Y^{\mathrm{m}}_{u_2}=\frac{\sum_{j=2}^{3} P_{\mathrm{r},1,j,u_2}+(1-\alpha)P_{\mathrm{r},1,1,u_2}}{\alpha P_{\mathrm{r},1,1,u}}. \tag{13}$$

In addition, let $I_{\mathrm{a},u}$ denote the aggregate interference from all neighbor sites to a UE u in sector $\mathcal{A}_{1,1}$, which is computed by

$$I_{\mathrm{a},u}=\sum_{i=2}^{N_{sites}} I_{i,u}. \tag{14}$$

Then, let $Y_{\mathrm{a},u}$ denote the ratio of the aggregate interference from neighbor sites to the signal of UE u (denoted by $\mathcal{S}_u$), which is computed by

$$Y_{\mathrm{a},u}=\frac{I_{\mathrm{a},u}}{\mathcal{S}_u}, \tag{15}$$

where, $\mathcal{S}_{u_1}=(1-\alpha)P_{\mathrm{r},1,1,u_1}\psi_{1,u_1}$ and $\mathcal{S}_{u_2}=\alpha P_{\mathrm{r},1,1,u_2}\psi_{1,u_2}$.

Based on the above new variables, the SIRs of u_1 and u_2 can be expressed as

$$\gamma^{\mathrm{m}}_{u_1,r}(u_1,u_2,\alpha)=\frac{1}{Y^{\mathrm{m}}_{u_1}+Y_{\mathrm{a},u_1}}, \tag{16}$$

$$\gamma^{\mathrm{m}}_{u_2,r}(u_1,u_2,\alpha)=\frac{1}{Y^{\mathrm{m}}_{u_2}+Y_{\mathrm{a},u_2}}. \tag{17}$$

It is noted that both $Y^{\mathrm{m}}_{u_2}$ and $Y^{\mathrm{m}}_{u_2}$ are deterministic variables, while Y_{a,u_1} and Y_{a,u_2} are random variables. To make the statistical analytical model tractable, the aggregate correlated interference $I_{\mathrm{a},u}$ of the neighbor sites to UE u is approximated by a log-normal variable, which offers a good accuracy in [35]. Let $\mu_{I_{\mathrm{a},u}}$ and $\sigma_{I_{\mathrm{a},u}}$ denote the mean and standard deviation of the normal distribution associated with the log-normal approximation $I_{\mathrm{a},u}$, respectively, which can be computed by a low complexity method presented in [35].

As $\mathcal{S}_u$ is a log-normal variable, with the approximation of $I_{\mathrm{a},u}$ as a log-normal variable, $Y_{\mathrm{a},u}$ in the form of (15) is known to be a log-normal variable as well. Let $\mu_{Y_{\mathrm{a},u}}$ and $\sigma_{Y_{\mathrm{a},u}}$ denote the mean and standard deviation of the normal distribution associated with the lognormal variable $Y_{\mathrm{a},u}$, respectively [35].

Then, $\mu_{Y_{\mathrm{a},u}}$ and $\sigma_{Y_{\mathrm{a},u}}$ can be computed by

$$\mu_{Y_{\mathrm{a},u}}=\mu_{I_{\mathrm{a},u}}-\ln(P_{\mathrm{r},1,1,u}), \tag{18}$$

$$\sigma^2_{Y_{\mathrm{a},u}}=\sigma^2_{I_{\mathrm{a},u}}+\sigma^2_{\mathrm{w}}-2\rho_{\mathrm{a},u}\sigma_{I_{\mathrm{a},u}}\sigma_{\mathrm{w}}. \tag{19}$$

After the mean $\mu_{Y_{\mathrm{a},u}}$ and standard deviation $\sigma_{Y_{\mathrm{a},u}}$ for log-normal variable $Y_{\mathrm{a},u}$ have been computed, the probability density functions of $\gamma^{\mathrm{m}}_{u_1,r}(u_1,u_2,\alpha)$ and $\gamma^{\mathrm{m}}_{u_2,r}(u_1,u_2,\alpha)$ given by (17) are determined accordingly. The above analysis is also applicable to the SIR distribution $\gamma^{\mathrm{o}}_{u,r}$ for a general UE with OMA. Let Y^{o}_u be an intra-site interference related variable for OMA, which is defined as

$$Y^{\mathrm{o}}_u=\frac{\sum_{j=2}^{3} P_{\mathrm{r},1,j,u_1}}{P_{\mathrm{r},1,1,u}}. \tag{20}$$

SIR $\gamma^{\mathrm{o}}_{u,r}$ can be expressed as a function of a log-normal variable $Y_{\mathrm{a},u}$, or

$$\gamma^{\mathrm{o}}_{u,r}=\frac{1}{Y^{\mathrm{o}}_u+Y_{\mathrm{a},u}}. \tag{21}$$

C. Spectral Efficiency for a Pair of UEs with Fixed Locations

For a general log-normal distributed random variable X, with its parameters μ and σ being the mean and standard deviation of X's natural logarithm, the probability density function (denoted by $f_X(x;\mu,\sigma)$) of X can be expressed as

$$f_X(x;\mu,\sigma)=\frac{1}{x\sigma\sqrt{2\pi}}\mathrm{e}^{-\frac{(lnx-\mu)^2}{2\sigma^2}}. \tag{22}$$

Let $\mathcal{F}(x)$ define a function to calculate the SE from UE SIR x. The instantaneous SEs of UEs, u_1 and u_2, with OMA, denoted by $C^{\mathrm{o}}_{u_1}$ and $C^{\mathrm{o}}_{u_2}$, respectively, can be computed by

$$C^{\mathrm{o}}_{u_1}=\mathcal{F}(\gamma^{\mathrm{o}}_{u_1,r}), \tag{23}$$

$$C^{\mathrm{o}}_{u_2}=\mathcal{F}(\gamma^{\mathrm{o}}_{u_2,r}). \tag{24}$$

The SE of two multiplexed UEs, u_1 and u_2, with NOMA under the condition of $\gamma^{\mathrm{o}}_{u_1,r}>\gamma^{\mathrm{o}}_{u_2,r}$, denoted by $C^{\mathrm{m}}_{u_1}(u_1,u_2,\alpha)$ and $C^{\mathrm{m}}_{u_2}(u_1,u_2,\alpha)$, respectively, can be computed by

$$C^{\mathrm{m}}_{u_1}(u_1,u_2,\alpha)=\mathcal{F}[\gamma^{\mathrm{m}}_{u_1,r}(u_1,u_2,\alpha)], \tag{25}$$

$$C^{\mathrm{m}}_{u_2}(u_1,u_2,\alpha)=\mathcal{F}[\gamma^{\mathrm{m}}_{u_2,r}(u_1,u_2,\alpha)]. \tag{26}$$

Let $C^{\mathrm{m}}_{\mathrm{sum}}(u_1,u_2,\alpha)$ represent the sum SE of two multiplexed UEs, u_1 and u_2, under the condition of $\gamma^{\mathrm{o}}_{u_1,r}>\gamma^{\mathrm{o}}_{u_2,r}$. In a NOMA system, a given pair of UEs with different locations are multiplexed if $C^{\mathrm{m}}_{\mathrm{sum}}(u_1,u_2,\alpha)$ is larger than $(C^{\mathrm{o}}_{u_1}+C^{\mathrm{o}}_{u_1})/2$ under the condition of $\gamma^{\mathrm{o}}_{u_1,r}>\gamma^{\mathrm{o}}_{u_2,r}$; otherwise the two UEs are served by OMA.

Let $C^{\mathrm{n}}(u_1,u_2,\alpha)$ denote the sum SE of a NOMA system under the condition of $\gamma^{\mathrm{o}}_{u_1,r}>\gamma^{\mathrm{o}}_{u_2,r}$. We can obtain

$$C^{\mathrm{n}}(u_1,u_2,\alpha)= \begin{cases} C^{\mathrm{m}}_{\mathrm{sum}}(u_1,u_2,\alpha), & \text{if } C^{\mathrm{m}}_{\mathrm{sum}}(u_1,u_2,\alpha)>(C^{\mathrm{o}}_{u_1}+C^{\mathrm{o}}_{u_1})/2,\\ (C^{\mathrm{o}}_{u_1}+C^{\mathrm{o}}_{u_1})/2, & \text{otherwise.}\end{cases} \tag{27}$$

As the SIRs of UEs, u_1 and u_2, are random variables, we compute the mean sum SE of the scheduled pair of UEs

in a NOMA system, which is denoted by $\overline{C^{\mathrm{n}}}(u_1,u_2,\alpha)$ and computed by

$$\overline{C^{\mathrm{n}}}(u_1,u_2,\alpha)=\int_{y_1+Y_{\mathrm{thr}}}^{\infty}\int_{0}^{\infty}C^{\mathrm{n}}(u_1,u_2,\alpha) f_{Y_{\mathrm{a},u_1}}(y_1;\mu_{Y_{\mathrm{a},u_1}},\sigma_{Y_{\mathrm{a},u_1}})f_{Y_{\mathrm{a},u_2}}(y_2;\mu_{Y_{\mathrm{a},u_2}},\sigma_{Y_{\mathrm{a},u_2}})\mathrm{d}y_1\mathrm{d}y_2 + \int_{0}^{y_1+Y_{\mathrm{thr}}}\int_{0}^{\infty}C^{\mathrm{n}}(u_2,u_1,\alpha) f_{Y_{\mathrm{a},u_1}}(y_1;\mu_{Y_{\mathrm{a},u_1}},\sigma_{Y_{\mathrm{a},u_1}})f_{Y_{\mathrm{a},u_2}}(y_2;\mu_{Y_{\mathrm{a},u_2}},\sigma_{Y_{\mathrm{a},u_2}})\mathrm{d}y_1\mathrm{d}y_2, \tag{28}$$

where $Y_{\mathrm{thr}}=Y^{\mathrm{o}}_{u_1}-Y^{\mathrm{o}}_{u_2}$, corresponding to the condition $\gamma^{\mathrm{o}}_{u_1,r}>\gamma^{\mathrm{o}}_{u_2,r}$. This condition is equivalent to $\frac{1}{Y^{\mathrm{o}}_{u_1}+Y_{\mathrm{a},u_1}}>\frac{1}{Y^{\mathrm{o}}_{u_2}+Y_{\mathrm{a},u_2}}$, which gives $Y_{\mathrm{a},u_2}>Y_{\mathrm{a},u_1}+Y^{\mathrm{o}}_{u_1}-Y^{\mathrm{o}}_{u_2}$. It is noted that both $C^{\mathrm{n}}(u_1,u_2,\alpha)$ and $C^{\mathrm{n}}(u_2,u_1,\alpha)$ are the functions of the integrands y_1 and y_2, which can be found from formulae (23) to (28). $\overline{C^{\mathrm{n}}}(u_1,u_2,\alpha)$ can be obtained by simple numerical integration tools.

Let $\overline{C^{\mathrm{n}}_{u_1}}(u_1,u_2,\alpha)$ and $\overline{C^{\mathrm{n}}_{u_2}}(u_1,u_2,\alpha)$ denote the mean SE of individual UEs, u_1 and u_2, in a scheduled pair with PAC α in a NOMA system. They can be computed using a similar formula derived earlier for $\overline{C^{\mathrm{n}}}(u_1,u_2,\alpha)$, which is not repeated here.

D. Numerical Results

1) Network Throughput: With the above analysis on the mean SE of a fixed pair of UEs in a NOMA system, network throughput and fairness performance can be modeled. Network SIR and outage probability can be analyzed in a similar way. To facilitate the network level performance analysis, the whole service area of sector $\mathcal{A}_{1,1}$ is divided into the segments with an equal size of $d_{\mathrm{res}}\times d_{\mathrm{res}}$ m^2. The segments in sector $\mathcal{A}_{1,1}$ are labeled from 1 to N_{g}, where N_{g} denotes the number of segments. Each segment has one and only one UE at its center. Let Ω_{seg} denote the set of UEs.

In the NOMA systems with RR scheduling, each UE has an equal probability to pair with any UE (including the UE itself) and to be scheduled. For any pair of scheduled UEs with different locations, the mean SE has been derived in the previous subsection. If a UE u is selected to pair with itself, the mean SE for this specific pair is denoted by $\overline{C^{\mathrm{o}}_u}$, which is computed by

$$\overline{C^{\mathrm{o}}_u}=\int_{0}^{\infty}C^{\mathrm{o}}_u f_{Y_{\mathrm{a},u}}(y;\mu_{Y_{\mathrm{a},u}},\sigma_{Y_{\mathrm{a},u}})\mathrm{d}y, \tag{29}$$

where C^{o}_u is the instantaneous SE of UE u.

Let $\overline{C^{\mathrm{n}}_u}(\alpha)$ denote the mean SE of a general UE u in a NOMA system for $u\in\Omega_{\mathrm{seg}}$, which is computed by averaging over the mean SE of u with all possible NOMA pairs, or

$$\overline{C^{\mathrm{n}}_u}(\alpha)=\frac{\sum_{u_1\in\Omega_{\mathrm{seg}},u_1\neq u}\overline{C^{\mathrm{n}}}(u,u_1,\alpha)+\overline{C^{\mathrm{o}}_{u_1}}}{N_{\mathrm{g}}}. \tag{30}$$

Let $\overline{C^{\mathrm{n}}_{\mathrm{net}}}(\alpha)$ and $\eta_{\mathrm{site}}(\alpha)$ denote the mean network SE and site throughput. We can compute $\overline{C^{\mathrm{n}}_{\mathrm{net}}}(\alpha)$ by

$$\overline{C^{\mathrm{n}}_{\mathrm{net}}}(\alpha)=\frac{\sum_{u\in\Omega_{\mathrm{seg}}}\overline{C^{\mathrm{n}}_u}(\alpha)}{N_{\mathrm{g}}}. \tag{31}$$

Accordingly, mean site throughput $\eta_{\mathrm{site}}(\alpha)$ can be computed by

$$\eta_{\mathrm{site}}(\alpha)=3B_{\mathrm{net}}\overline{C^{\mathrm{n}}_{\mathrm{net}}}(\alpha), \tag{32}$$

where B_{net} denotes network bandwidth in Hz, and the site throughput is computed from the three sectors of the site. UE throughput fairness, denoted by $F_{\mathrm{net}}(\alpha)$, is computed with Jain's fairness index formula as

$$F_{\mathrm{net}}(\alpha)=\frac{\Big[\sum_{u\in\Omega_{\mathrm{seg}}}\overline{C^{\mathrm{n}}_u}(\alpha)\Big]^2}{N_{\mathrm{g}}\sum_{u\in\Omega_{\mathrm{seg}}}\overline{C^{\mathrm{n}}_u}(\alpha)^2}. \tag{33}$$

2) Numerical Results for NOMA with RR Scheduling: Next, representative numerical results for NOMA systems with RR scheduler are presented, which were obtained with an analytical model and system level simulations. System configuration is shown in Table II. In the simulations, UEs are randomly and uniformly distributed in a network. The PAC α varies from 0.55 to 1 with a step size of 0.05. 10,000 snapshots are captured to compute UE and network statistics. For each snapshot, shadowing fading and small scale fading coefficients for the sectors and UEs are generated. Each UE is paired with UEs in the network and all the UE pairs are scheduled in one snapshot. Sum SE for the multiplexing pairs in the NOMA system is computed with various PAC values. Shannon channel formula is used, but the network performance model is applicable to other spectral efficiency models.

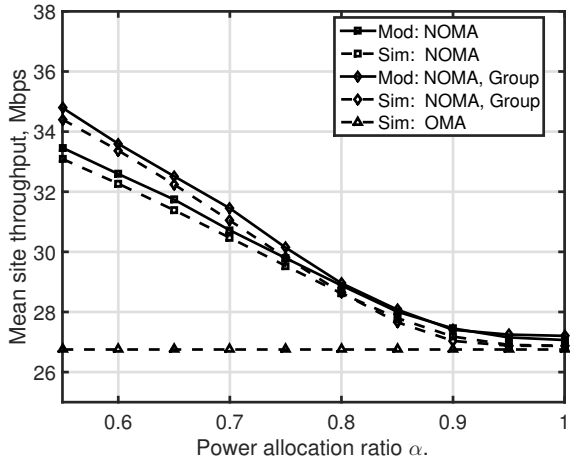
In addition to the OMA and basic NOMA systems, the performance of a group based NOMA system (denoted by Group-NOMA) is also presented. In the literature, it was suggested that a UE should be multiplexed with another UE when its channel quality is better than a given threshold in order to increase NOMA performance [4], [19]. To test how effective such an approach can be, UEs are divided into two UE groups according to their mean SIRs, and each UE is only allowed to pair up with the UEs in the opposite group. The analysis on the joint RR scheduling for Group-NOMA system is similar to that for the basic NOMA system. Figs. 2(a) and 2(b) present the throughput and fairness index against α, respectively. It can be observed that the analytical results match to system-level simulation results closely. In addition, the system level simulations took much more time than the analytical approach.

According to Fig. 2(a), the site throughput of basic NOMA and Group-NOMA systems decreases almost linearly with α, which confirms the analysis of NOMA power allocation in Section III-A. The throughput of an OMA system is not affected by α. The throughput gain of basic NOMA over OMA systems is 23.7% and 10.4% with α=0.55 and 0.75, respectively. The Group-NOMA system has additional 5%

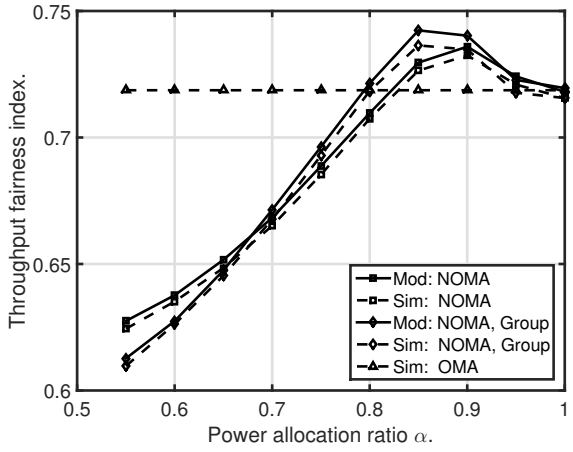
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2018.2869305, IEEE Transactions on Vehicular Technology

throughput gain over the basic NOMA system at α=0.55, but the gain diminishes with an increasing α. Therefore, setting multiplexing threshold does not give a substantial performance improvement. According to Fig. 2(b), the UE fairness of both basic NOMA and Group-NOMA systems improves with α until α reaches 0.85, and then degrades gradually to the value of an OMA system.

With the fast and accurate analytical model, the NOMA systems with RR scheduling can be evaluated effectively, and a utility function can be set up to control the NOMA parameter α, considering both network throughput and fairness. A recommended range for α is 0.65 to 0.75.



(a) Mean site throughput (Mbps).



(b) UE throughput fairness.

Fig. 2. a) Mean site throughput; b) UE throughput fairness of a NOMA system with RR scheduling.

IV. CROSS-LAYER DESIGN OF NOMA WITH PROPORTIONAL FAIRNESS SCHEDULING

In this section, the issues on cross-layer design and optimization of NOMA and PF scheduling algorithms will be investigated. While RR and maximum signal to interference ratio schedulers are designed to maximize UE fairness and network throughput, a PF scheduler maintains a good balance between total network throughput and UE fairness. PRBs are assigned to UEs according to a scheduling priority, which is inversely proportional to anticipated resource consumption of the UEs. Due to the attractive features of PF scheduling, it has been widely used as a reference scheduler in many works on NOMA [2]–[4]. However, as PF scheduling has a much higher computational complexity, and PF scheduling and NOMA are coupled much more tightly than the RR scheduling, it is crucial to keep the computational complexity of joint scheduling with NOMA algorithms as low as possible.

Next, the weighted sum rate and optimization problem for joint PF scheduling with NOMA algorithms are studied in Section IV-A. A two-step approach is proposed for joint design of NOMA and PF scheduling algorithms.

A. Weighted Sum Rate with NOMA and Exhaustive Search Algorithm

In LTE cellular networks, time is split up into TTIs with index t. To facilitate the introduction of PF scheduling algorithm, we add a subscript t to the previous introduced variable notations, such as SIR and SE of UEs with OMA and NOMA. For example, now we use $\gamma^{\rm o}_{u,r,t}$ and $C^{\rm o}_{u,r,t}$ to denote the SIR and SE of UE u with OMA over PRB r at TTI t. NOMA SIRs for two multiplexed UEs, u_1 and u_2, with PAC α are denoted by $\gamma^{\rm m}_{u_1,r,t}(u_1,u_2,\alpha)$ and $\gamma^{\rm m}_{u_2,r,t}(u_1,u_2,\alpha)$ over PRB r at TTI t. Let $\overline{\eta}_{u,t}$ be the average throughput of UE u in the last W TTIs. W is known as the time window of PF scheduling, which is set to 50.

Resource allocation is centrally controlled by eNBs, which assign different PRBs to different UEs depending on so-called scheduling priority coefficient (PC), which is calculated as a function of UE channel states and their average throughput. Let $\varphi_{u,r,t}$ denote the PC of UE u over PRB r at TTI t with OMA, which is computed by

$$\varphi_{u,r,t} = \frac{B_{\rm rb} C^{\rm o}_{u,r,t}}{\overline{\eta}_{u,t}}. \tag{34}$$

In the OMA systems with PF scheduling, in a general TTI (say t), eNB computes the PC of all UEs over all PRBs according to (34). Initially, all PRBs are added to a resource set $\Omega_{\rm rb}$. The PRBs are allocated to UEs one unit per round. In the first round of PRB allocation, PRB r^* in the set $\Omega_{\rm rb}$ with the maximal PC is allocated to UE u^* with the largest PC over PRB r^*. The allocated PRB r^* is removed from the resource set $\Omega_{\rm rb}$. And UE average throughput and its PCs over the PRBs in the set $\Omega_{\rm rb}$ are updated. The above process is repeated until all PRBs are allocated.

In the NOMA systems with PF scheduling, resource allocation becomes more challenging, as each PRB may be allocated to more than one UE and there can be multiple power allocation levels for UE multiplexing. For each PRB allocation, eNB needs to choose only one UE or a group of multiplexed UEs, a proper α for multiplexing, and resource allocation optimization objectives. In the existing works studying joint PF scheduling for NOMA, a widely accepted optimization objective for PRB allocation is the weighted sum rate of multiplexed UEs over PRB [2]. Let $\varphi_{u_1,u_2,r,t}(\alpha)$ denote the

TABLE II
SYSTEM PARAMETER SETTINGS

Parameters	Value	Parameters	Value
Carrier frequency	2000 MHz	Bandwidth	B_{net}=5 MHz; B_{rb}=180 KHz
Number of sites	$N_{\text{site}} = 19$	Inter-site distance	500 m
Transmit power	21.6 W	Shadowing	σ_{w} = 6 dB; $\rho = 0.5$
Antenna height	25 m	Max antenna gain (dBi)	15.5
Antenna front to back ratio	25 dB	HPBW	horizontal: 65°; vertical: 11.5°
UE density (simulation)	0.0025 per m^2	Segment resolution	d_{res} = 10 m

weighted sum rate of UEs, u_1 and u_2, which are multiplexed over PRB r at TTI t. It is computed by

$$\varphi_{u_1,u_2,r,t}(\alpha) = \frac{B_{\text{rb}} C^{\text{m}}_{u_1,r,t}(u_1,u_2,\alpha)}{\overline{\eta}_{u_1,t}} + \frac{B_{\text{rb}} C^{\text{m}}_{u_2,r,t}(u_1,u_2,\alpha)}{\overline{\eta}_{u_2,t}}, \tag{35}$$

which can be easily extended to the cases where more than two UEs are multiplexed over a PRB.

For the joint PF scheduling in NOMA systems, the operation of PF scheduling and NOMA UPPA is tightly coupled. For each PRB allocation, the UEs to be selected depend on not only UEs' own channel state but also the sum SE of UEs when they are multiplexed with various power allocation options. To maximize the weighted sum rate over all N_{rb} PRBs for a general TTI with index t, an optimization problem can be formulated as

$$\begin{aligned} \max_{u_1,u_2,\alpha} \quad & \sum_{r=1}^{N_{\text{rb}}} \varphi_{u_1,u_2,r,t}(\alpha), \\ \text{subject to:} \quad & u_1, u_2 \in \Omega_{\text{ue}}, \\ & \gamma^{\text{m}}_{u,r,t}(u_1,u_2,\alpha) > \gamma_{\min},\ u \in \Omega_{\text{ue}}, \\ & \alpha \in [0,1]. \end{aligned} \tag{36}$$

For the decision making on the user selection, user pairing and power allocation, a straightforward algorithm is exhaustive search (ES) algorithm [2]. With the ES algorithm, the weighted sum rate for all the combinations of UE groups sharing a PRB and power allocation options are computed first. The UE groups and the power allocation with the largest weighted sum rate over the PRB is chosen to get the PRB. It is well known that such an ES algorithm has a very high computational complexity. Therefore, fast joint PF scheduling and NOMA UPPA algorithms are needed.

B. Optimal Power Allocation for a Pair of UEs

In this subsection, we introduce an optimal power allocation strategy for any given pair of UEs (say u_1 and u_2), multiplexed over a general PRB at a TTI. The optimization objective is the weighted sum rate $\varphi_{u_1,u_2,r,t}(\alpha)$ computed by (35). Note that in Section III-A, optimal power allocation for a given pair of multiplexed UEs was analyzed with optimization objective of sum SE. It was shown that the optimal strategy is to simply allocate full transmit power to a UE with a better channel quality.

For the analysis on optimal power allocation with an objective of maximizing weighted sum rate, we use the Shannon capacity formula. It is noted that such an assumption is only used for the optimal power allocation analysis. In the design of UE pairing algorithms and simulations, different channel capacity formulae can be used.

As the analysis is applicable to a general TTI index and PRB, subscripts t and r are neglected in the variables used in this subsection. Eqn. (35) for weighted sum rate of UEs, u_1 and u_2, is rewritten as

$$\begin{aligned} \varphi_{u_1,u_2}(\alpha) &= \frac{B_{\text{rb}} C^{\text{m}}_{u_1}(u_1,u_2,\alpha)}{\overline{\eta}_{u_1}} + \frac{B_{\text{rb}} C^{\text{m}}_{u_2}(u_1,u_2,\alpha)}{\overline{\eta}_{u_2}} \\ &= \frac{B_{\text{rb}}}{\overline{\eta}_{u_2}} \log_2 \left\{ \left[1+\gamma^{\text{m}}_{u_1}(u_1,u_2,\alpha)\right]^d \left[1+\gamma^{\text{m}}_{u_2}(u_1,u_2,\alpha)\right] \right\} \\ &= \frac{B_{\text{rb}}}{\overline{\eta}_{u_2}} \log_2 \left[(1+\beta\gamma^{\text{o}}_{u_1})^d \frac{1+\gamma^{\text{o}}_{u_2}}{1+\beta\gamma^{\text{o}}_{u_2}} \right] \\ &= \frac{B_{\text{rb}}}{\overline{\eta}_{u_2}} \left\{ \log_2 \left[\frac{(1+\beta\gamma^{\text{o}}_{u_1})^d}{1+\beta\gamma^{\text{o}}_{u_2}} \right] + \log_2(1+\gamma^{\text{o}}_{u_2}) \right\}, \end{aligned} \tag{37}$$

where $d = \frac{\overline{\eta}_{u_2}}{\overline{\eta}_{u_1}}$ and $\beta = 1-\alpha$ with $0 \le \beta < 0.5$.

From (37) it is known that maximizing $\varphi_{u_1,u_2}(\alpha)$ is equivalent to maximizing the factor $\frac{(1+\beta\gamma^{\text{o}}_{u_1})^d}{1+\beta\gamma^{\text{o}}_{u_2}}$. Differentiating the factor against β, we get

$$\begin{aligned} &\frac{\text{d}\frac{(1+\beta\gamma^{\text{o}}_{u_1})^d}{1+\beta\gamma^{\text{o}}_{u_2}}}{\text{d}\beta} \\ &= \frac{(1+\beta\gamma^{\text{o}}_{u_1})^{d-1}}{(1+\beta\gamma^{\text{o}}_{u_1})^2} \left[d\gamma^{\text{o}}_{u_1}(1+\beta\gamma^{\text{o}}_{u_2}) - \gamma^{\text{o}}_{u_2}(1+\beta\gamma^{\text{o}}_{u_1}) \right]. \end{aligned} \tag{38}$$

It is noted that the sign of the above differentiation in (38) depends solely on the second factor.

Checking Eqn. (38), we can develop the following optimal power allocation strategy for a given pair of multiplexed UEs as follows.

- If $d > 1$ or $d\gamma^{\text{o}}_{u_1} - \gamma^{\text{o}}_{u_2} < 0.5\gamma^{\text{o}}_{u_1}\gamma^{\text{o}}_{u_2}(1-d)$, we always have $d\gamma^{\text{o}}_{u_1}(1+\beta\gamma^{\text{o}}_{u_2}) - \gamma^{\text{o}}_{u_2}(1+\beta\gamma^{\text{o}}_{u_1}) > 0$. Therefore, the weighted sum rate φ_{u_1,u_2} increases monotonically with β, and thus β (and α as well) should take a value close to 0.5.
- If $d\gamma^{\text{o}}_{u_1} < \gamma^{\text{o}}_{u_2}$, φ_{u_1,u_2} decreases monotonically with β, we set α to 1 to maximize φ_{u_1,u_2}.
- Otherwise, set α to $1 - \frac{d\gamma^{\text{o}}_{u_1} - \gamma^{\text{o}}_{u_2}}{\gamma^{\text{o}}_{u_1}\gamma^{\text{o}}_{u_2}(1-d)}$ to maximize φ_{u_1,u_2}.

C. Joint PF Scheduling and UE Pairing with $N_m = 2$

After the analysis of optimal power allocation problem, the joint PF scheduling and UE pairing problem is investigated. In [28] the authors proposed a simple joint PF scheduling and UE pairing algorithm, which is called PF-Fast algorithm.

In the PF-Fast algorithm, instead of comparing all UE pairs to search for the best UE pair optimizing weighted sum rate for a PRB, one UE (which is selected as the one having the largest OMA weighted rate over the PRB) from each pair is compared in the PF-Fast algorithm. This identified UE is called the first multiplexing UE (MUE) of a multiplexing pair. The pairs formed by the first MUE with all UEs are compared (including the first MUE itself). The second MUE of the pair to get the PRB is identified in the pair maximizing the weighted sum rate over the PRB. It is noted that the second MUE can be the same as the first MUE, which corresponds to OMA.

Simulation results presented in [28] show that the PF-Fast algorithm can reduce computation time without a big performance loss. However, the PF-Fast algorithm still has a high computational complexity of $\mathcal{O}(N_{\text{ue}}N_{\text{rb}})$. And the PF-Fast algorithm is only applicable to the cases with at most two UEs multiplexing. Next, three joint PF scheduling and UPPA algorithms are proposed to reduce computational complexity. The new algorithms are faster and more scalable (applicable to the case that a large number of UEs share a PRB). We present the new algorithms for the case of at most two UEs sharing a PRB (N_{m} =2). Then, we extend the algorithms to the cases with imperfect CSI and with more UEs sharing a PRB.

1) PF-FS-SIR and PF-FS-PC Algorithms: In the new joint scheduling and use pairing algorithms, the first MUE of the pairs sharing unallocated PRBs is identified with the same approach used in the PF-Fast algorithm [28]. The difference lies on the way how to choose the second MUE of multiplexed pair. Instead of comparing the pairs formed by all UEs in the network with the first MUE, only a few UEs are selected as the candidates to pair with the first MUE and are compared for possible PRB allocation.

There are several criteria that can be used to select the UEs to pair up with the first MUE. An algorithm using SIR criteria to choose the second MUE is proposed first, which is called PF-FS-SIR algorithm. The letters "FS" in the algorithm's name refer to being fast and scalable. In the PF-FS-SIR algorithm, only two UEs are chosen to pair up with the first MUE and compared in terms of the weighted sum rate, i.e., the first MUE itself (corresponding to OMA, a specific case of NOMA with α=0), and then the UE with the largest OMA SIR. If the UE with the largest OMA SIR is different from the first MUE, α for NOMA multiplexing of this UE pair can be set to a pre-configured value, such as 0.75, or determined by the optimal power allocation method. The overall PF-FS-SIR algorithm with optimal power allocation at TTI t is presented in **Algorithm** 1.

In the PF-FS-SIR algorithm, pairing the first MUE having the largest OMA weighted rate with the second MUE having the largest SIR may deliver a higher network throughput but poorer UE throughput fairness. The UEs with good channel states have more chances to be multiplexed and receive a larger share of frequency resources. To address the fairness problem, a new algorithm named as PF-FS-PC is proposed. The idea is to pair the first MUE having the largest OMA weighted rate with the second MUE having the second largest OMA weighted rate. The above multiplexing pair is compared to the first MUE in terms of weighted sum rate to decide if OMA or NOMA multiplexing should be applied over a PRB. PF-FS-PC algorithm takes a very similar approach as PF-FS-SIR, as shown in **Algorithm** 1.

Algorithm 1 Joint PF scheduling and UE pairing PF-FS-SIR.

Input: Weighted rate $\varphi_{u,r,t}$ for $u \in \Omega_{\text{ue}}$, $r \in \Omega_{\text{rb}}$.
Output: UEs scheduled over each PRB r, $r \in \Omega_{\text{rb}}$.
Initialize the set of unallocated PRB Ω_{rb} to $\{1, ..., N_{\text{rb}}\}$.
while $\Omega_{\text{rb}} \neq \emptyset$ **do**
  Find PRB r^* to be processed next and first MUE u_1^*:
  $(r^*, u_1^*) = \underset{u\in\Omega_{\text{ue}}, r\in\Omega_{\text{rb}}, \gamma^{\text{o}}_{u,r}>\gamma_{\min}}{\arg\max} \varphi_{u,r,t}$
  Find u_2^*: $u_2^* = \underset{u\in\Omega_{\text{ue}}\setminus\{u_1^*\}, \gamma^{\text{o}}_{u,r^*}>\gamma_{\min}}{\arg\max} \gamma^{\text{o}}_{u,r^*,t}$
  Compute $\varphi_{u_1^*,u_2^*,r^*,t}$ with either fixed or optimal power allocation according to (35)
  if $\varphi_{u_1^*,r^*,t} > \varphi_{u_1^*,u_2^*,r^*,t}$ **or** $\gamma^{\text{m}}_{u_1^*,r}(u_1^*, u_2^*, \alpha) < \gamma_{\min}$ **or** $\gamma^{\text{m}}_{u_2^*,r}(u_1^*, u_2^*, \alpha) < \gamma_{\min}$ **then**
    Replace u_2^* by u_1^*; using OMA.
  end if
  Allocate PRB r^* to UEs u_1^* and u_2^*
  Update Ω_{rb} by $\Omega_{\text{rb}} \leftarrow \Omega_{\text{rb}} \setminus \{r^*\}$
  Update $\overline{\eta}_{u_1^*,t}$ and $\overline{\eta}_{u_2^*,t}$
  Update weighted rate $\varphi_{u_1^*,r,t}$ and $\varphi_{u_2^*,r,t}$, for all $r \in \Omega_{\text{rb}}$.
end while

2) PF-FS-Hybrid Algorithm: It is noted that the objective of the PF-FS-PC algorithm is aligned with the original PF scheduling objective, that is to maximize PC for PRBs to be allocated. But a potential issue of the PF-FS-PC algorithm is that simply multiplexing the UEs with the largest and the second largest OMA weighted rate may lead to a lower instantaneous UE throughput and mean network throughput. In view of the strength and weakness of PF-FS-SIR and PF-FS-PC algorithms, a new algorithm named as PF-FS-Hybrid is proposed. In the PF-FS-Hybrid algorithm, as usual, the PRB to be allocated to the first MUE having the largest OMA weighted rate over that PRB are identified. Then, three multiplexing pairs formed by the next second MUEs with the first MUE are compared, i.e., the first MUE itself, a different MUE with the highest OMA SIR, and another different MUE with the second highest OMA weighted rate. The PRB is allocated to one of the above three multiplexing pairs having the largest weighted sum rate.

Note that if one of the second MUE candidates, such as the one with the largest OMA SIR (or OMA weighted rate), is excluded, PF-FS-Hybrid algorithm behaves exactly the same as PF-FS-SIR (or PF-FS-PC) algorithm.

D. Joint Algorithm with $N_m = 2$ and Imperfect CSI

In the previous subsections, perfect CSI was assumed in the development of the joint scheduling and NOMA UPPA algorithms. However, due to fast fading channel state estimation is not always perfect. It is important to investigate how imperfect CSI estimation will affect the performance of joint scheduling and NOMA algorithms and how to improve the design of joint algorithms with imperfect CSI. For simplicity,

we assume a simple channel estimation model. Let $\gamma_{u,r}^{\mathrm{o}}$ and $\gamma_{u,r}^{\mathrm{o,e}}$ denote the estimated and actual channel SIRs of a user u over PRB r with OMA, respectively. The estimated SIR $\gamma_{u,r}^{\mathrm{o}}$ is available to the eNB. The actual $\gamma_{u,r}^{\mathrm{o,e}}$ is unknown to eNB, which is expressed by

$$\gamma_{u,r}^{\mathrm{o,e}} = (1 + e_{u,r})\gamma_{u,r}^{\mathrm{o}}, \tag{39}$$

where $e_{u,r}$ represents the estimation error on the estimated SINR $\gamma_{u,r}^{\mathrm{o}}$. Similar to the assumption made in [24], $e_{u,r}$ is assumed to follow a Gaussian distribution with mean zero and variance σ_{e}^2. Let $f_{\mathcal{N}}(x)$ denote the probability distribution function (PDF) of the Gaussian distribution. It is noted that alternative distributions for CSI estimation error can be used. A better distribution of the estimation error and/or distribution parameters can be obtained through analysis on channel estimation samples.

In this subsection, we extend the PF-FS-Hybrid algorithm, knowing that the CSI estimation is imperfect. We first present the computation of the mean SIR and spectrum efficiency at eNB with actual CSI, which are used in the extended PF-FS-Hybrild algorithm with imperfect CSI. Then, we present the changes made to the PF-FS-Hybrid algorithm designed with perfect CSI.

Let $C_{u,r}^{\mathrm{o,e}}$ denote the mean spectrum efficiency of UE u over PRB r with OMA and CSI estimation error, where e designates imperfect CSI. Assuming the use of Shannon capacity formula, we have

$$C_{u,r}^{\mathrm{o,e}} = \int_{-1}^{\infty} \log_2[1 + (1 + x)\gamma_{u,r}^{\mathrm{o}}] f_{\mathcal{N}}(x)\mathrm{d}x. \tag{40}$$

The outage probability that the actual SIR of UE u over r with OMA is smaller than γ_{min} can be computed by

$$\mathrm{Prob}(\gamma_{u,r}^{\mathrm{o,e}} < \gamma_{\mathrm{min}}) = \int_{-\infty}^{\frac{\gamma_{\mathrm{min}} - \gamma_{u,r}}{\gamma_{u,r}}} f_{\mathcal{N}}(x)\mathrm{d}x. \tag{41}$$

Consider multiplexing u_1 and u_2 over PRB r with NOMA. Let $\gamma_{u_1,r}^{\mathrm{m,e}}(u_1, u_2, \alpha)$ and $\gamma_{u_2,r}^{\mathrm{m,e}}(u_1, u_2, \alpha)$ denote the actual SIR of u_1 and u_2 over r with NOMA. The actual SIR of multiplexed UEs, u_1 and u_2, can be expressed by their estimated SIR with OMA, or

$$\gamma_{u_1,r}^{\mathrm{m,e}}(u_1, u_2, \alpha) = (1 + e_{u_1,r})(1 - \alpha)\gamma_{u_1,r}^{\mathrm{o}}, \tag{42}$$

$$\gamma_{u_2,r}^{\mathrm{m,e}}(u_1, u_2, \alpha) = \frac{1 + (1 + e_{u_2,r})\gamma_{u_2,r}^{\mathrm{o}}}{1 + (1 + e_{u_2,r})(1 - \alpha)\gamma_{u_2,r}^{\mathrm{o}}}. \tag{43}$$

With the actual SIR of UEs, u_1 and u_2 multiplexed with NOMA, the outage probabilities of u_1 and u_2 can be computed and used in the UPPA decision making. Let $C_{u_1,r}^{\mathrm{m,e}}(u_1, u_2, \alpha)$ and $C_{u_2,r}^{\mathrm{m,e}}(u_1, u_2, \alpha)$ denote the mean spectrum efficiencies of u_1 and u_2 been multiplexed over r with NOMA and with CSI estimation error. From (42) and (42), we have

$$\begin{aligned} &C_{u_1,r}^{\mathrm{m,e}}(u_1, u_2, \alpha) \\ &= \int_{-1}^{\infty} \log_2[1 + (1 + x)(1 - \alpha)\gamma_{u_1,r}^{\mathrm{o}}] f_{\mathcal{N}}(x)\mathrm{d}x, \end{aligned} \tag{44}$$

and

$$\begin{aligned} &C_{u_2,r}^{\mathrm{m,e}}(u_1, u_2, \alpha) \\ &= \int_{-1}^{\infty} \log_2 \left[1 + \frac{1 + (1 + x)\gamma_{u_2,r}^{\mathrm{o}}}{1 + (1 + x)(1 - \alpha)\gamma_{u_2,r}^{\mathrm{o}}}\right] f_{\mathcal{N}}(x)\mathrm{d}x. \end{aligned} \tag{45}$$

From (40) for OMA spectrum efficiency $C_{u,r}^{\mathrm{o,e}}$, and (44) and (45) for NOMA spectrum efficiency, the OMA weighted rate and NOMA weighted rate under imperfect CSI can be computed in a similar way as (34) and (35), respectively, by replacing the estimated spectrum efficiency with the mean spectrum efficiency.

Using the outage probabilities, updated SIR and weighted rate with imperfect CSI, we can modify the PF-FS-Hybrid algorithm for the NOMA systems with imperfect CSI. It is noted that PF-FS-Hybrid algorithm is chosen for the investigation with imperfect CSI due to its good overall performance. Other algorithms presented in the previous subsection can be modified in a similar way. In the new algorithm, there are three major changes made against the original PF-FS-Hybrid algorithm, or

- For the QoS constraint on SIR, any UE u allocated to a PRB r (through either OMA or NOMA) has to satisfy a revised condition that the outage probability is lower than the outage probability threshold O_{min}.
- The mean UE weighted rate and the mean SIR with actual CSI are used in the process of selecting the best UEs for OMA and NOMA, instead of the instantaneous weighted rate and SIR.
- The weighted sum of mean rates with actual CSI is used in the selection of OMA or NOMA and the UE pairing.

E. Joint Algorithm with $N_m > 2$

So far, only the case for at most two UEs sharing a PRB is discussed. As allowing more than two UEs multiplexing over a PRB can yield a larger performance gain, PF-FS-Hybrid algorithm is extended to work with $N_{\mathrm{m}} > 2$. A similar extension can be easily made to PF-FS-SIR and PF-FS-PC algorithms. Similar to the design of the joint PF scheduling and UPPA algorithms with $N_{\mathrm{m}} = 2$, the design problem with $N_{\mathrm{m}} > 2$ can be decomposed into two sub-problems, i.e., a) power allocation for a given number of UEs expected to share a PRB; b) PF scheduling and UE grouping.

1) Power Allocation Sub-problem: First, let us consider the power allocation sub-problem. Assume that there are n UEs (say u_1, u_2, ..., and u_n) to be multiplexed over a PRB. Without loss of generality, the OMA SIRs of these UEs are arranged in an increasing order. In the case of $n = 1$, all transmit power P_{t} is allocated to u_1. In the case of $n = 2$, transmit powers αP_{t} and $(1 - \alpha)P_{\mathrm{t}}$ are allocated to u_1 and u_2, respectively. The coefficient α can be determined by the optimal power allocation strategy or using a pre-configured value. For $n > 2$, we propose a simple iterative method to allocate the transmit power. The power allocation starts from the multiplexing of the first two UEs (u_1 and u_2). The transmit power $\alpha_1 P_{\mathrm{t}}$ allocated to u_1 is determined first, where α_1 is found using the same approach for the case of two UE multiplexing. For the multiplexing of UEs u_2 and u_3, they share the remaining transmit power $(1-\alpha_1)P_{\mathrm{t}}$. Transmit powers $(1-\alpha_1)\alpha_2 P_{\mathrm{t}}$ and $(1-\alpha_1)(1-\alpha_2)P_{\mathrm{t}}$ are allocated to u_2 and u_3, respectively. The coefficient α_2 is determined again using the power allocation approach for two UEs multiplexing. The above process is repeated to allocate transmit power for the remaining UEs.

At the end of the power allocation process for $N_{\rm m} > 2$, the following formula is obtained for the transmission power $P_{{\rm t},u_l}$ allocated to UE u_l, for $l = 1, ..., n$, or

$$P_{{\rm t},u_l} = \begin{cases} \alpha_1 P_{\rm t}, & l = 1, \\ \prod_{i=1}^{l-1}(1-\alpha_i)\alpha_l P_{\rm t}, & 1 < l < n, \\ \prod_{i=1}^{n-1}(1-\alpha_i) P_{\rm t}, & l = n. \end{cases} \quad (46)$$

2) Scheduling and UE Pairing Sub-problem: Next, the joint scheduling and UE pairing sub-problem is investigated. The above power allocation strategy is embedded in the joint scheduling and UE pairing algorithm as presented earlier, which is used to compute SIR when more than one UE is multiplexed. To facilitate the design of the PF-FS-Hybrid algorithm with $N_{\rm m} > 2$, a binary tree based NOMA multiplexing graph is introduced. At the first level $n = 0$, the binary tree has only one node (the root), which corresponds to the first MUE (with the largest OMA weighted rate over a considered PRB). Note that a level in the binary tree means the distance to the root. At the second level n=1, two children (i.e., two second MUEs) are added to the root (the first MUE). The second MUEs with the largest SIR and the largest weighted rate among the UEs not presented in the higher level of the binary tree are added as the left and right children of the root, respectively. The above process repeats until level $n = N_{\rm m}-1$. An example multiplexing tree with $N_{\rm m} = 4$ is shown in Fig. 3. It is noted that $u_{n,sir}$ and $u_{n,pc}$ represent the MUEs with the largest SIR and PC at tree level $n-1$, respectively, for $n = 0, ..., N_{\rm m} - 1$.

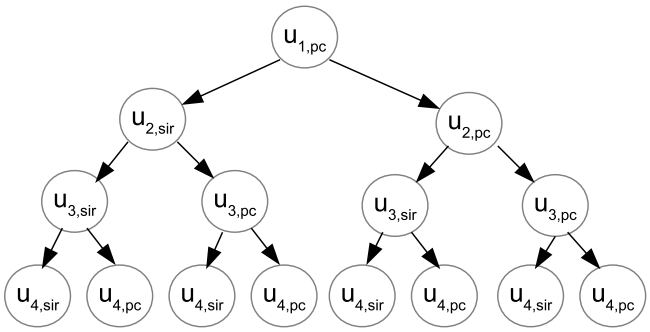


Fig. 3. An example multiplexing tree with $N_{\rm m} = 4$, where $u_{n,sir}$ and $u_{n,pc}$ represent the MUEs with the largest SIR and PC at tree level $n-1$, respectively.

In the PF-FS-Hybrid algorithm, for each PRB allocation, the candidate groups of UEs sharing a PRB can be represented by all the paths starting from the root $u_{1,\rm pc}$ to all the nodes in the multiplexing binary tree. For example, at tree level 0, we have one candidate UE for PRB allocation, which is $u_{1,\rm pc}$. At tree level 1, there are two nodes and two paths from the root, representing two candidate groups of UEs for possible sharing the PRB, $\{u_{1,\rm pc}, u_{2,\rm sir}\}$ and $\{u_{1,\rm pc}, u_{2,\rm pc}\}$. And at tree level 2, there are four paths from the root: $\{u_{1,\rm pc}, u_{2,\rm sir}, u_{3,\rm sir}\}$, $\{u_{1,\rm pc}, u_{2,\rm sir}, u_{3,\rm pc}\}$, $\{u_{1,\rm pc}, u_{2,\rm pc}, u_{3,\rm sir}\}$, and $\{u_{1,\rm pc}, u_{2,\rm pc}, u_{3,\rm pc}\}$. After finding all the candidate groups of UEs, the weighted sum rates of the multiplexed UEs in the candidate groups over a PRB are compared, and the PRB is allocated to the UE group with the largest weighted sum rate.

For computational complexity analysis, we introduce a basic unit of computation, *SNM*, to compute the weighted sum SE for a pair of NOMA multiplexed UEs. In the ES based PF scheduling and NOMA algorithm, the total number of candidate UE multiplexing pairs is $\frac{N_{\rm ue}^2}{2}$. Taking the number of power allocation levels into account, we get that the total number of *SNM* computations is $\frac{N_{\rm pa}N_{\rm ue}^2}{2}$ for one PRB. In one TTI, eNB needs to perform $\frac{N_{\rm rb}N_{\rm pa}N_{\rm ue}^2}{2}$ *SNM* computations over $N_{\rm rb}$ PRBs. Thus, the computational complexity of PF-ES algorithm is $\mathcal{O}(N_{\rm rb}N_{\rm pa}N_{\rm ue}^2)$. The computation complexity for other algorithms can be analyzed similarly.

V. PERFORMANCE OF JOINT PF SCHEDULING AND NOMA ALGORITHMS

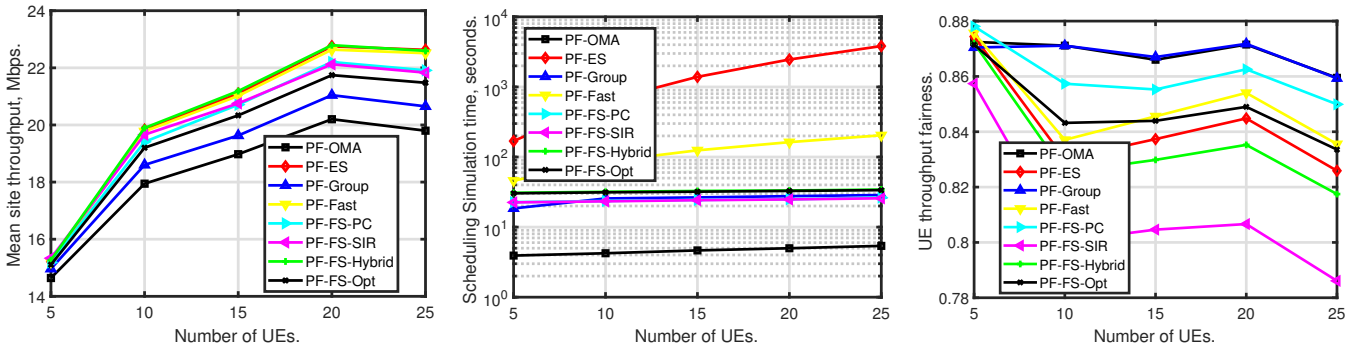
In this section, the proposed PF scheduling and NOMA algorithms are compared to the existing algorithms. The system configuration is shown in Table II. As theoretic analysis of PF scheduling is very difficult, system level simulations are used for evaluation of NOMA systems with PF scheduling. For each algorithm, five drops of UEs to the network are considered and 5,000 simulation snapshots are run for each UE drop to obtain mean network performance. Shadow fading coefficients between eNBs and UEs are generated once and used for all the simulations. 3GPP spatial channel model (SCM) is used to generate fast fading coefficients [3], [34]. The minimum SIR $\gamma_{\rm min}$ set for UE resource allocation is -2 dB. The outage probability threshold $O_{\rm min}$ is set to 0.1. The proposed algorithms are evaluated first with ideal CSI at eNBs and then with imperfect CSI and $N_{\rm m} > 2$.

A. Comparison of Various Scheduling and NOMA Algorithms with N_m=2

Eight scheduling and NOMA algorithm settings are evaluated and compared. The features of the evaluated algorithms are summarized in Table III. Although all NOMA based algorithms can use either fixed or optimal power allocation in NOMA multiplexing, only fixed power allocation with α=0.75 is used, except that PF-FS-Opt algorithm uses the optimal power allocation. Columns "1st MUE" and "2nd MUE" refer to how the first and second MUEs of UE pairs are found to share a PRB. PF-ES algorithm compares all UEs for both the first and second MUEs with a fixed α. The PF-ES algorithm with a full search over all power allocation levels is not evaluated, as it is too time-consuming for the simulations, and it is demonstrated that optimal power allocation does not give a better network performance than fixed power allocation.

One compared algorithm that is not introduced in Section IV is PF-Group algorithm. In this algorithm, UEs are allocated to two groups according to their OMA SIRs for each PRB allocation. The first MUE is chosen as the UE with the largest OMA PC over that PRB, and the second MUE is found as the UE with the largest OMA PC from the opposite group of the first MUE. In the original grouping based UE pairing algorithm [4], each UE in one group will form pairs with

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2018.2869305, IEEE Transactions on Vehicular Technology



(a) Site throughput with N_{rb}=8. (b) Schedule time with N_{rb}=8. (c) UE fairness with N_{rb}=8.

Fig. 4. Network performance against the number of UEs with N_{rb}=8. a) and b): Mean site throughput; c) and d): schedule time per drop; e) and f): UE throughput fairness.

TABLE III
PARAMETER SETTINGS OF COMPARED ALGORITHMS.

Name	N_{m}	α	1st MUE	2nd MUE	# of SNM
PF-OMA	1	0.75	-	-	0
PF-ES	2	0.75	All	All	$N_{\text{rb}}N_{\text{ue}}^2/2$
PF-Fast	2	0.75	PC	All	$N_{\text{rb}}N_{\text{ue}}$
PF-Group	2	0.75	Group PC	Group PC	N_{rb}
PF-FS-PC	2	0.75	PC	PC	$2N_{\text{rb}}$
PF-FS-SIR	2	0.75	PC	SIR	$2N_{\text{rb}}$
PF-FS-Hybrid	2	0.75	PC	SIR+PC	$3N_{\text{rb}}$
PF-FS-Opt	2	Optimal	PC	SIR+PC	$3N_{\text{rb}}$

every UE in the other group, and the UE pairs with the largest weighted sum rate are chosen to share the PRBs. In the original grouping algorithm, the number of *SNM* computations per PRB is $N_{\text{ue}}^2/4$, which is at the same as that of PF-ES algorithm. Therefore, the modified version of grouping based algorithm is used for comparison instead of the original version.

The performance metrics of interest include mean site throughput, simulation time spent on scheduling (and UPPA) per drop, and UE throughput fairness. Fig. 4 shows the mean site throughput, scheduling and UPPA time, and UE fairness versus the number of UEs for eight PRBs. Shannon channel formula is used to compute spectral efficiency. Simulation results with other spectral efficiency formulae show similar performance trends, which are not presented due to space limit.

From Fig. 4, the following are observed on the performance of the compared algorithms.

1) Throughput: The algorithms with NOMA consistently outperform PF-OMA. There are two clusters of PF scheduling and NOMA algorithms, in which algorithms deliver very close throughput. The first cluster has PF-ES, PF-Fast, and PF-FS-Hybrid algorithms. The second cluster has PF-FS-SIR, PF-FS-PC, and PF-FS-Opt algorithms. The algorithms in the first cluster have around 14% throughput gain over PF-OMA algorithm with 20 UEs. The algorithms in the second cluster have around 10% throughput gain. PF-Group algorithm is the second worst, with 5% throughput gain over PF-OMA in the 20 UEs scenario.
2) Simulation time on scheduling: PF-OMA takes the shortest scheduling time as expected. Scheduling simulation time of PF-OMA and the proposed fast algorithms increases very little with the number of UEs. On the other hand, PF-ES simulation time increases significantly with N_{ue}. In the scenario with N_{ue}=20, scheduling time of PF-ES is 60 times that of the proposed algorithms. It is noted that the overall simulations (including extra simulation tasks, such as channel state generation, SIR computation, and statistic collection, etc.) took a much longer time. In total, it took five days to complete the simulations to produce the set of results shown in Fig. 4, which indicates how time consuming the PF-ES algorithm is.
3) Fairness: Fairness is another important metric for NOMA algorithms. As shown in Fig. 4, PF-OMA, PF-Group, and PF-FS-PC have the best fairness as expected. PF-Fast and PF-FS-Opt come next. PF-ES and PF-FS-Hybrid are slightly worst than PF-FS-Opt. PF-FS-SIR has the worst fairness performance, which is expected as it always chooses the second MUE with the largest SIR.

The results demonstrated the complex interactions between PF scheduling and NOMA algorithms. For example, PF-FS-PC algorithm with fixed power allocation (α=0.75) has consistently better performance than PF-FS-Opt. It suggests that optimal power allocation with the objective of maximizing PF scheduling priority coefficient does not guarantee an optimal network performance (e.g., throughput and fairness). Therefore, optimal power allocation may not be needed for PF scheduling and NOMA. On the other hand, the performance of PF-FS-SIR algorithm, which attempts to maximize the network throughput by pairing the first MUE with the UE having the largest SIR, is consistently worst than PF-FS-Hybrid algorithm.

PF-FS-Hybrid algorithm shows a good overall performance, i.e., high throughput, low scheduling time, and comparable fairness to PF-ES algorithm. Apart from that, PF-Fast algorithm has identical throughput of PF-ES, but much lower scheduling time and better fairness.

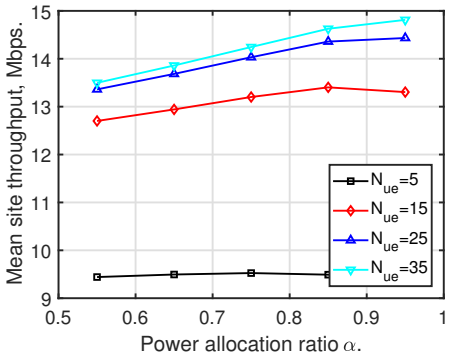
B. Impact of Power Allocation, imperfect CSI, and Maximal Number of Multiplexed UEs

In the algorithm comparison presented in the previous subsection, PAC α was fixed as 0.75, perfect CSI estimation

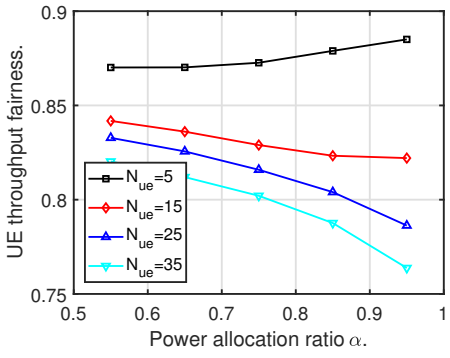
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2018.2869305, IEEE Transactions on Vehicular Technology

was assumed, and N_{m} was two. In this subsection, the impact of fixed power allocation with various PAC, imperfect CSI and larger N_{m} settings is examined. Only PF-FS-Hybrid algorithm is used for performance investigation in this subsection, as it is one of the best performing algorithms.

The network throughput and UE fairness against α are presented in Figs. 5(a) and 5(b), respectively. It can be observed clearly that network throughput increases with α until α reaches 0.85, but UE fairness reduces in most cases with an increasing α. The impact of α on NOMA with PF scheduling performance is quite different from that with NOMA and RR scheduling. Setting α to 0.75 strikes a good tradeoff on the network throughput and fairness.



(a) Site throughput.

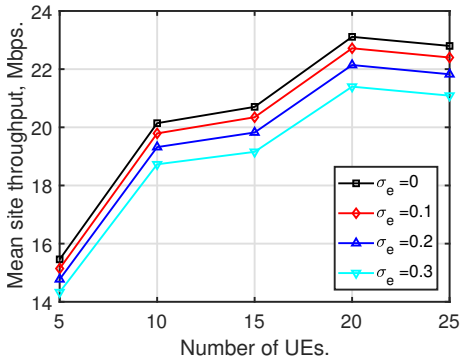


(b) UE fairness.

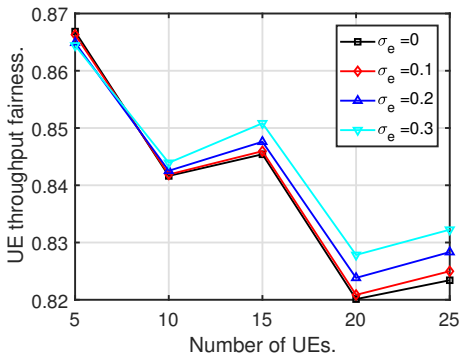
Fig. 5. PF-FS-Hybrid algorithm performance versus α with N_{rb}=5 and N_{m}=2. a) Mean site throughput; b) UE throughput fairness.

Next, the impact of imperfect CSI is studied. The standard deviation σ_{e} of the CSI estimation error is set to 0, 0.1, 0.2, and 0.3 in the experiments. Note that perfect CSI is a specific case in the study with $\sigma_{\mathrm{e}} = 0$. The network throughput and user fairness versus the number of UEs with eight PRBs and N_{m} =2 are presented in Figs. 6(a) and 6(b), respectively. The results show that the proposed algorithm is robust in the presence of CSI estimation error. With an increasing CSI estimation error, the overall network throughput drops gradually, while the user fairness increases slightly. Even with CSI estimation error $\sigma_{\mathrm{e}} = 0.3$, the network throughput of the PF-FS-Hybrid algorithm is still higher than that of PF-OMA algorithm with perfect CSI.

To investigate the impact of multiplexing more UEs, N_{m} is set to 2, 3, and 4. Mean site throughput and scheduling time of PF-FS-Hybrid algorithm are presented against the



(a) Site throughput.



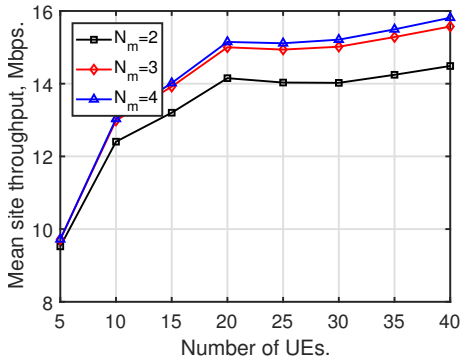
(b) Schedule time per drop.

Fig. 6. PF-FS-Hybrid algorithm performance versus N_{ue} with imperfect CSI, where N_{rb}=8 and α=0.75. a) Site throughput; b) UE throughput fairness.

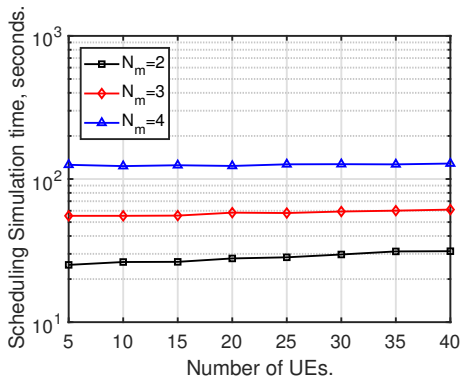
number of UEs in Figs. 7(a) and 7(b), respectively. α is set to 0.75. According to Fig. 7(a), increasing N_{m} from 2 to 3 leads to around 1 Mbps throughput improvement for scenarios of $N_{\mathrm{ue}} > 10$, corresponding to around 7% throughput gain. With the performance gain of PF-FS-Hybrid over PF-OMA at N_{m}=2, PF-FS-Hybrid has around 24% throughput gain over PF-OMA at N_{m}=3. But there is marginal further throughput gain by increasing N_{m} to four. According to Fig. 7(b), scheduling time does not change much with the number of UEs. Increasing N_{m} from 2 to 3 doubles the scheduling time, which is still acceptable. The scheduling time performance against the number of UEs and N_{m} demonstrates that the proposed algorithms have a low computational complexity and a good scalability. Considering both throughput gain and scheduling time, we know that N_{m}=3 is a good choice for NOMA multiplexing.

VI. CONCLUSION

In this paper, joint design and optimization of scheduling and UPPA algorithms for NOMA systems was investigated. First, the impact of power allocation for joint NOMA and round-robin scheduling algorithm was analyzed. Then, design and optimization solution for joint PF scheduling and NOMA algorithms with QoS constraints was developed with a two-step approach. In the first step, impact of power allocation was analyzed. An optimal algorithm was proposed for a given pair of multiplexed UEs with the objective of maximizing weighted sum of rate. Three fast scheduling and UE pairing algorithms with different design objectives on network throughput and



(a) Site throughput.



(b) Schedule time per drop.

Fig. 7. PF-FS-Hybrid algorithm performance versus N_{m} with N_{rb}=5 and α=0.75. a) Site throughput; b) Schedule time per drop.

fairness were designed. The proposed algorithms examine only one or two carefully selected UE pairs, which are formed with the UE having the largest weighted rate. The proposed scheduling and UPPA algorithms were extended to the NOMA systems with imperfect CSI and more than two UEs multiplexing. Simulation results validated the high speed and effectiveness of the proposed algorithms. The computational complexity of UPPA was significantly reduced.

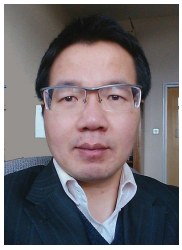
ACKNOWLEDGMENTS

The authors would like to thank the reviewers for their constructive comments. The work of J. He and Z. Tang was supported by FP7 grant DETERMINE (FP7-PEOPLE-2012-IRSES under grant number 318906) and EU Horizon2020 grant COSAFE (H2020-MSCA-RISE-2018 under grant number 824019).

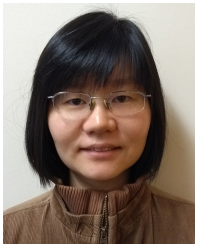
REFERENCES

[1] V. W. S. Wong, *et al.*, "Key Technologies for 5G Wireless Systems", Cambridge University Press, 2017.
[2] Y. Saito, Y. Kishiyama, A. Benjebbour, *et al.*, "Non-orthogonal multiple access (NOMA) for cellular future radio access," *IEEE VTC Spring'13*, Dresden, Germany, June 2013.
[3] Y. Saito, A. Benjebbour, *et al.*, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)," *in Proc. IEEE PIMRC'13*, pp. 611-615, 2013.
[4] A. Benjebbour, A. Li, Y. Saito, *et al.*, "System-level performance of downlink NOMA for future LET enhancements," *Globecom Workshops'13*, Atlanta, GA, USA, Dec. 2013.
[5] L. Dai, B. Wang, *et al.*, "Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74-81, Sep. 2015.
[6] H. Nikopour, E. Yi, *et al.*, "SCMA for downlink multiple access of 5G wireless networks," *in Proc. GLOBECOM*, pp. 1-5, Dec. 2014.
[7] S. Islam, N. Avazov, *et al.*, "Power-Domain Non-Orthogonal Multiple Access (NOMA) in 5G Systems: Potentials and Challenges," *IEEE Commun. Surveys and Tutorials*, Oct. 2016.
[8] Z. Ding, *et al*, "A Survey on Non-Orthogonal Multiple Access for 5G Networks: Research Challenges and Future Trends," *IEEE Journal of Selected Areas on Commun.*, vol. 35, no. 10, pp. 2181 - 2195, Oct. 2017.
[9] H. Zhang, *et al.*, "Energy-Efficient Resource Allocation in NOMA Heterogeneous Networks," IEEE Wireless Communications, 2018 (In Press).
[10] D. N. Tse, "Optimal power allocation over parallel Gaussian broadcast channels," *Proceedings of IEEE International Symposium on Information Theory*, Ulm, pp. 27, 1997.
[11] Lifang Li and A. J. Goldsmith, "Capacity and optimal resource allocation for fading broadcast channels .I. Ergodic capacity," *IEEE Transactions on Information Theory*, vol. 47, no. 3, pp. 1083-1102, Mar 2001.
[12] D. Tse and P. Viswanath, "Fundamentals of Wireless Communication," *Cambridge University Press*, 2005.
[13] Z. Ding, Z. Yang, *et al.*, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1501-1505, Dec. 2014.
[14] Z. Ding and H. Poor, "Design of Massive-MIMO-NOMA With Limited Feedback," *IEEE Signal Processing Lett.*, vol. 23, no. 5, pp. 629-633, May 2016.
[15] Z. Ding, F. Adachi, V. Poor, *et al.*, "The application of MIMO to Non-orthogonal multiple access," *IEEE Trans. Wireless Comm.*, **15**, (1), p.537-552, Jan. 2016.
[16] W. Shin, M. Vaezi, *et al.*, "Coordinated Beamforming for Multi-Cell MIMO-NOMA," *IEEE Commun. Lett.*, vol. 21, no. 1, pp. 84-87, Jan. 2017.
[17] J. Men and J. Ge, "Non-orthogonal multiple access for multiple-antenna relaying networks," *IEEE Commun. Lett.*, vol. 19, no. 10, pp. 1686-1689, Oct. 2015.
[18] L. Lei, D. Yuan, *et al.*, "Power and Channel Allocation for Non-Orthogonal Multiple Access in 5G Systems: Tractability and Computation," *IEEE Trans. on Wireless Commun.*, vol. 15, no. 12, pp. 8580-8594, Dec. 2016.
[19] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G non-orthogonal multiple access downlink transmissions," *IEEE Trans. on Veh. Technol.*, vol. 65, no. 8, pp. 6010 - 6023, Aug. 2016.
[20] L. Qian, Y. Wu, H. Zhou, and X. Shen, "Non-Orthogonal Multiple Access Vehicular Small Cell Networks: Architecture and Solution," *IEEE Network*, pp. 15-21, July 2017.
[21] P. Parida and S. S. Das, "Power allocation in OFDM based NOMA systems: A DC programming approach," *Proc. IEEE Globecom Workshops*, Austin, TX, pp. 1026-1031, 2014.
[22] S. Liu, C. Zhang, and G. Lyu, "User selection and power schedule for downlink non-orthogonal multiple access (NOMA) system," *in Proc. IEE ICCW'15*, pp. 2561-2565, 2015.
[23] F. Fang, *et al.*, "Energy-Efficient Resource Allocation for Downlink Non-Orthogonal Multiple Access Network," *IEEE Transactions on Communications*, vol. 64, no. 9, pp. 3722-3732, Sept. 2016.
[24] F. Fang, *et al.*, "Joint User Scheduling and Power Allocation Optimization for Energy Efficient NOMA Systems With Imperfect CSI," IEEE Journal on Selected Areas in Commun., vol. 35, no. 12, pp. 2874-2885, Dec. 2017.
[25] Y. Sun, *et al.*, "Optimal Joint Power and Subcarrier Allocation for Full-Duplex Multicarrier Non-Orthogonal Multiple Access Systems," *IEEE Trans. on Commun.*, vol. 65, no. 3, pp. 1077-1091, March 2017.
[26] Z. Wei, *et al.*, "Optimal Resource Allocation for Power-Efficient MC-NOMA With Imperfect Channel State Information," *IEEE Trans. on Commun.*, vol. 65, no. 9, pp. 3944-3961, Sept. 2017.
[27] Jianyue Zhu, *et al.*, "On Optimal Power Allocation for Downlink Non-Orthogonal Multiple Access Systems", *IEEE Journal on Selected Areas in Commun.*, vol. 35, pp. 2744-2757, Dec. 2017.
[28] J. He, Z. Tang, and Z. Che, "Fast and efficient UE pairing and power allocation algorithm for non-orthogonal multiple access in cellular networks," *Electronics Letters*, vol. 52, no. 25, pp. 2065-2067, Dec. 2016.
[29] J. He and Z. Tang, "Low-complexity user pairing and power allocation algorithm for 5G cellular network non-orthogonal multiple access," *Electronics Letters*, vol. 53, no. 9, pp. 626-627, Apr. 2017.
[30] B. Di, L. Song, and Y. Li, "Sub-channel Assignment, Power Allocation, and User Scheduling for Non-orthogonal Multiple Access Networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7686-7698, Nov. 2016.
[31] J. Choi, "On HARQ-IR for Downlink NOMA Systems," *IEEE Trans. Commun.*, vol. 64, no. 8, pp. 3576-3584, Aug. 2016.

[32] J. Choi, "Power Allocation for Max-Sum Rate and Max-Min Rate Proportional Fairness in NOMA," *IEEE Commun. Lett.*, vol. 20, no. 10, pp. 2055-2058, Oct. 2016.
[33] S. Schwartz and Y. Yeh, "On the distribution function and moments of power sums with lognormal components," Bell System Technology Journal, Vol. 61, No. 7, pp. 1441-1462, Sept. 1982.
[34] 3GPP TR 36.814 V9.0.0, "Further advancements for E-UTRA physical layer aspects," *Technical Report*, March 2010.
[35] J. He, W. Cheng, Z. Tang, *et al*, "Analytical evaluation of higher order sectorization, frequency reuse, and UE classification methods in OFDMA networks," *IEEE Trans. Wireless Comm.*, vol. 15, no. 12, pp. 8209-8222, Dec. 2016.
[36] A. Ligeti, "Outage probability in the presence of correlated lognormal useful and interfering components," *IEEE Commun. Lett.*, Vol. 4, No. 1, pp. 15-17, Jan. 2000.



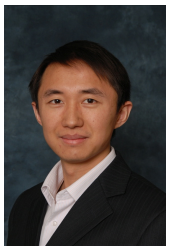
Jianhua He received his BSc and MSc degrees in Electronic Information Engineering from Huazhong University of Science and Technology (HUST), China, and a PhD degree from Nanyang Technological University, Singapore, in 1995, 1998 and 2002, respectively. He joined HUST in 2001 as an Associate Professor. From 2004 to 2011, he has been with University of Bristol, University of Essex and University of Swansea. Dr He is a Lecturer at Aston University, UK. His main research interests are 5G networks, connected vehicles, autonomous driving, Internet of things, AI for OCR and wireless newtorks. He has authored or co-authored over 100 technical papers in major international journals and conferences. Dr He is the coordinator of EU Horizon2020 project COSAFE on connected autonomous vehicles. He is an IEEE Senior Member.



Zuoyin Tang is currently a Lecturer in the School of Engineering and Applied Science, Aston University, UK. She obtained her PhD degree from University of Bath, UK, in 2008. She has authored and co-authored over 40 technical papers in major international journals and conferences. Dr Tang's main research interests include resource management for cellular networks, Internet of things and wireless sensor networks.

Zuowen Tang is currently a PhD student in the School of Engineering and Applied Science, Aston University, UK. She obtained her BSc degree in Electronic Information Enginneering from Wuhan University, China, in 2006, and MPhil degree from Reading University, UK, in 2011. Her main research interests include scheduling algorithms and resource management for cellular networks.

Hsiao-Hwa Chen (S'89-M'91-SM'00-F'10) is currently a Distinguished Professor in the Department of Engineering Science, National Cheng Kung University, Taiwan. He obtained his BSc and MSc degrees from Zhejiang University, China, and a PhD degree from the University of Oulu, Finland, in 1982, 1985 and 1991, respectively. He is the founding Editor-in-Chief of Wiley?s Security and Communication Networks Journal (http://www.interscience.wiley.com/security). Prof. Chen is an active volunteer for IEEE over 35 years. He is the recipient of 2016 IEEE Jack Neubauer Memorial Award. He served as the Editor-in-Chief for IEEE Wireless Communications from 2012 to 2015, and was an elected Member at Large of IEEE ComSoc from 2014 to 2016. He is a Fellow of IEEE, and a Fellow of IET. Currently, he is also a ComSoc Distinguished Speaker.



Cong Ling received the B.S. and M.S. degrees in electrical engineering from the Nanjing Institute of Communications Engineering, Nanjing, China, in 1995 and 1997, respectively, and the PhD degree in electrical engineering from the Nanyang Technological University, Singapore, in 2005. He is currently a Reader in the Electrical and Electronic Engineering Department at Imperial College London. His research interests are coding, information theory, and security, with a focus on lattices. Dr. Ling has served as an Associate Editor of IEEE Transactions on Communications and of IEEE Transactions on Vehicular Technology.