




Delay Minimization for Massive Internet of Things With Non-Orthogonal Multiple Access

Daosen Zhai , *Member, IEEE*, Ruonan Zhang , *Member, IEEE*, Lin Cai , *Senior Member, IEEE*, and F. Richard Yu , *Fellow, IEEE*

Abstract—Non-Orthogonal Multiple Access (NOMA) provides potential solutions for the stringent requirements of the Internet of Things (IoT) on low latency and high reliability. In this paper, we jointly consider user scheduling and power control to investigate the access delay minimization problem (ADMP) for the uplink NOMA networks with massive IoT devices. Specifically, the ADMP is formulated as a mixed-integer and non-convex programming problem with the objective to minimize the maximum access delay of all devices under individual data transmission demand. We prove that the ADMP is NP-hard. To tackle this hard problem, we divide it into two subproblems, i.e., the user scheduling subproblem (USP) and the power control subproblem (PCP), and then propose an efficient algorithm to solve them in an iterative manner. In particular, the USP is recast as a K-CUT problem and solved by a graph-based method. For the PCP, we devise an iterative algorithm to solve it optimally leveraging the standard interference function. Simulation results indicate that our algorithm has good convergence and can significantly reduce the access delay in comparison with other schemes.

Index Terms—Graph theory, internet of things, non-orthogonal multiple access, resource management.

I. INTRODUCTION

AS AN important component of the fifth generation (5G) mobile communication systems, the Internet of Things (IoT) has received rapid development in recent years. Tens of billions of new devices (e.g., wearable devices, smart appliances, autonomous vehicles, etc.) will access to the wireless networks. The forecast shows that there will be 26 billion connections by 2020 [1], and this number will increase fivefold in the following decade [2]. Massive connectivity is a key feature of the IoT. On the other hand, some special applications of the IoT, such as the industry control and automatic drive, re-

quire millisecond end-to-end latency [2]. The new requirements of the IoT on ultra-low latency and ultra-high connectivity pose serious challenges for 5G wireless networks. To meet these challenges, advanced enabling wireless technologies are needed.

Non-orthogonal multiple access (NOMA) is a promising technique, which can accommodate multiple users on the same spectrum by utilizing different power levels [3]. Besides, theoretical analysis indicates that NOMA can achieve a larger capacity region in comparison with the traditional orthogonal multiple access (OMA) techniques [4], e.g., time-division multiple access (TDMA) and orthogonal frequency-division multiple access (OFDMA). Thanks to the potential advantages in the network capacity, NOMA has been suggested to be an important multiple access technique in the future 5G wireless networks [5]. Furthermore, since more users can be supported by the NOMA networks simultaneously, the transmission delay of users can also be reduced significantly. Therefore, NOMA provides potential solutions to meet the rigorous requirements of the IoT on ultra-low latency and ultra-high connectivity [6]. The NOMA-based cellular architecture has been proposed to support a massive number of IoT devices in cellular networks [7].

However, NOMA also induces intra-cell interference among users, which complicates the decoding algorithm and degrades the individual data rate. To deal with this problem, extensive researches have been conducted on resource management in NOMA networks [8]–[35]. The control policies include power control, channel assignment, user clustering, user scheduling, user association, rate control, etc. Through appropriate resource management, the intra-cell interference among NOMA users is coordinated and the user diversities in the power domain are exploited, so the performance of NOMA networks can be upgraded. From the perspective of optimization goals, the researches fall into two major categories: 1) rate improvement works [8]–[18] and 2) energy conservation works [19]–[28]. The details of the related works are introduced in the following.

For the first category, the works aim to improve the individual data rate or the throughput of the whole network. Specifically, a dynamic power allocation scheme was proposed in [8] to enhance the average rate of each user in both uplink and downlink NOMA systems. With the same control policy, the author in [9] focused on the MIMO-NOMA network and optimized the transmission power on different layers of users to maximize the sum-rate of the network. In [10], an optimal and a low-complexity near-optimal joint power allocation and scheduling solutions were designed to ensure proportional fair resource al-

Manuscript received September 11, 2018; revised January 12, 2019; accepted February 1, 2019. Date of publication February 11, 2019; date of current version May 22, 2019. This work was supported in part by China Postdoctoral Science Foundation under Grants BX20180262 and 2018M641019, in part by the National Natural Science Foundation of China under Grants 61571370 and 61601365, in part by the Fundamental Research Funds for the Central Universities under Grant 3102017OQD091, and in part by the Civil Aircraft Major Project of China under Grant MIZ-2015-F-009. (*Corresponding author: Ruonan Zhang.*)

D. Zhai and R. Zhang are with the School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: zhaidaosen@nwpu.edu.cn; rzhang@nwpu.edu.cn).

L. Cai is with the Department of Electrical and Computer Engineering, University of Victoria, Victoria, BC V8P 5C2 Canada (e-mail: cai@ece.uvic.ca).

F. R. Yu is with the Department of Systems and Computer Engineering, Carleton University, Ottawa, ON K1S 5B6, Canada (e-mail: richard.yu@carleton.ca).

Digital Object Identifier 10.1109/JSTSP.2019.2898643

location and utility maximization in a simple NOMA system deploying two-layer hierarchical modulation. The authors in [11] investigated how to schedule multiple users into n given slots so as to maximize the throughput of the system or maintain certain level of fairness among users. Taking channel allocation into account, the authors in [12] and [13] designed the joint power and channel allocation algorithms for single-cell and heterogeneous networks, respectively. As a common feature, both [12] and [13] adopt the many-to-many matching method to solve the channel allocation problem. Different from [12] and [13], the authors in [14] formulated the channel allocation problem as a maximum weighted independent set problem and solved it by a graph-based method. The tractability and computation features of the joint power and channel allocation problem in NOMA networks were studied in [15]. Besides, based on the Lyapunov optimization framework, an online algorithm with rate control and power allocation was devised in [16] to maximize the long-term network utility. By jointly optimizing the user clustering and power allocation, the sum-throughput maximization problem was investigated in [17] for both uplink and downlink NOMA transmissions. In [18], the authors proposed a NOMA based secure transmission scheme, where the relay node not only received the privacy information but also provided securely information forwarding for the destination [36].

Regarding the other category, the works focus on minimizing the total transmission power or maximizing the network energy efficiency. In detail, the optimization of beamforming vectors and powers was studied in [19], with the objective to minimize the total transmission power of the base station (BS). To minimize the long-term average power consumption of the whole system (i.e., BS and devices), the authors in [20] proposed a dynamic user scheduling and power allocation algorithm based on the stochastic optimization theory. Besides, the distributed power control algorithm was proposed in [21] to minimize the total power consumption. Focusing on the multi-carrier networks, [22] and [23] investigated the joint channel and power allocation problem, where the goal of [22] was to minimize the total transmission power, while [23] concentrated in the energy efficiency improvement. Different from [19]–[23], the works in [24] and [25] considered the imperfect channel state information (CSI). More specifically, the joint rate and power allocation problem was studied in [24] to minimize the power consumption with throughput constraints. The authors in [25] investigated the energy efficiency improvement for a downlink NOMA network by jointly considering the user scheduling and power allocation. The works in [19], [20], [22]–[25] focused on the single-cell network. Different from them, the works in [26]–[28] aimed at the heterogeneous networks (HeNets). In [26], the traffic offloading scheme was designed for NOMA based HeNets with dual-connectivity. The authors in [27] investigated the user association and power control schemes in energy cooperation enabled two-tier HetNets for maximizing the energy efficiency of the overall network. In [28], the energy efficiency resource allocation algorithm was designed for user-centric ultra-dense networks, where the macro cells and micro cells can cooperate to transmit data through NOMA.

In addition to the above two categories, there are also some works investigating the connectivity [29], [30], outage probability [31]–[33], and fairness [34], [35] problems. The existing works have improved the performance of NOMA networks significantly. Nevertheless, there are still some problems remained to be further investigated. Specifically, there are very few works paying attention to the delay performance of NOMA networks. However, as indicated in the white paper on 5G versions and requirements [2], low latency is an important performance indicator of 5G wireless networks, e.g., the millisecond-level end-to-end latency and one millisecond air interface latency. Furthermore, some special applications of the IoT (e.g., the industry control and automatic drive) are more susceptible to the communication delay, as a small increment of delay may result in a tremendous loss. The recent work in [37] studied the offloading delay minimization problem for two paired NOMA users in a mobile edge computing system. Two iterative algorithms have been developed to optimize the transmission power of the paired NOMA users so as to minimize the offloading delay. However, the algorithms proposed in [37] are not suitable for the networks with massive connections. When the network is with massive connections, concurrent transmissions by all users are almost impossible even with NOMA. In this case, how to schedule the transmissions to minimize the access delay becomes an important problem. To deal with this problem, new resource management scheme should be carefully designed.

Motivated by the above, we investigate the delay minimization problem for NOMA networks with massive connections. The main contributions of this paper are summarized as follows.

- We jointly consider user scheduling and power control to investigate the access delay minimization problem (ADMP) in uplink NOMA networks with massive IoT devices. Specifically, the ADMP is formulated as a mixed-integer and non-convex programming problem, with the objective to minimize the maximum access delay of all devices under individual data transmission demand. Besides, we prove that the formulated problem is NP-hard.
- We propose an efficient algorithm to solve the formulated problem, which solves the user scheduling subproblem (USP) and the power control subproblem (PCP) in an iterative manner. In particular, the USP is remodeled as a K-CUT problem in graph theory, and then we devise a low-complexity algorithm to solve it. Afterward, the lower bound of the user scheduling algorithm is analyzed. For the PCP, we transform it into a min-max power control problem and solve it optimally based on the standard interference function.
- We conduct extensive simulations to evaluate the performance of our proposed algorithms. Firstly, the convergence performance of our algorithms is investigated. Furthermore, the scheduling results of different algorithms are presented to identify why our algorithm can achieve better performance. Moreover, the simulation results demonstrate that our algorithm can reduce almost 80% access delay with respect to the random NOMA scheme and the orthogonal multiple access scheme.

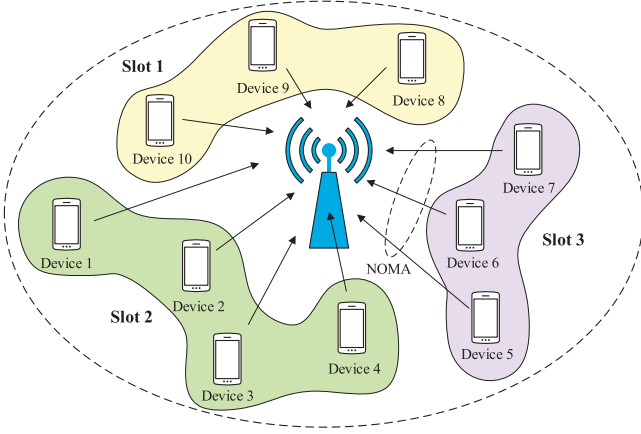


Fig. 1. The scenario of an uplink NOMA network with massive connections.

The remainder of this paper is organized as follows. In Section II, we introduce the network model and the problem formulation. In Section III, we analyze the characteristics of the formulated problem. Section IV elaborates the proposed user scheduling algorithm and power control algorithm. Simulation results are presented in Section V. Finally, we conclude our paper in Section VI.

II. NETWORK MODEL AND PROBLEM FORMULATION

In this section, we first introduce the considered network model and then present the problem formulation.

A. Network Model

As depicted in Fig. 1, we consider a NOMA uplink network, which consists of an access point (AP) and a massive number of IoT devices (e.g., sensor nodes, wearable devices, monitors, etc.). The set of the devices is denoted as $\mathcal{U} = \{1, 2, \dots, M\}$, where M represents the number of devices. We define G_m as the channel power gain (CPG) from the device m to the AP. For simplicity of expression, the devices are sorted in the descending order of their CPG (i.e., $G_m \geq G_n, \forall m < n$). It is assumed that the CPG is static or quasi-static in an operating period, that is, G_m remains constant in an operating period. Each operating period consists of T slots, denoted by $\mathcal{T} = \{1, 2, \dots, T\}$. Without loss of generality, the duration of a slot is normalized to one.

In this network, the IoT devices can collect information from surrounding environment and deliver them to the AP periodically. The typical applications include environmental sensing, remote health monitoring, and intelligent transportation, etc. The data harvested by device m is denoted as D_m , which must be transmitted to the AP during an operating period. To support massive connections, NOMA is adopted as the multiple access technique of this network. Specifically, the devices can transmit data to the AP on the same spectrum simultaneously, and the AP can separate the overlapping signals by the successive interference cancellation (SIC) technique. Although NOMA improves the spectrum efficiency, scheduling all of the devices in the same slot induces serious intra-cell interference, which coun-

teracts the performance gain of NOMA. In order to coordinate the intra-cell interference and fully exploit the advantages of NOMA, appropriate user scheduling and power control scheme should be carefully designed.

Specifically, we define $\mathbf{S}(t) = \{s_m(t) \mid m \in \mathcal{U}\}$ as the user scheduling policy in slot t , where $s_m(t) = 1$ if device m is scheduled in slot t , otherwise $s_m(t) = 0$. Besides, we denote \mathbf{S} as the user scheduling policies in an operational period, i.e., $\mathbf{S} = \{\mathbf{S}(t) \mid t \in \mathcal{T}\}$. The power control policy is defined as $\mathbf{P} = \{p_m \mid m \in \mathcal{U}\}$, where p_m represents the transmission power of device m . According to [4], [38], the optimal decoding order of SIC in uplink NOMA transmission should be the descending order in the CPG. As a consequence, the devices with weaker CPG cause interference to the devices with stronger CPG, but not vice versa. Therefore, the signal-to-interference-plus-noise ratio (SINR) of device m in slot t can be expressed as

$$\gamma_m(t) = \frac{s_m(t) p_m G_m}{\sum_{n \in \mathcal{U}, n > m} s_n(t) p_n G_n + \sigma^2}, \quad (1)$$

where σ^2 denotes the additive white Gaussian noise.

According to the Shannon Capacity theory, the achievable data rate of device m in slot t is

$$R_m(t) = B_0 \log_2(1 + \gamma_m(t)), \quad (2)$$

where B_0 is the channel bandwidth.

The total data volume transmitted by device m in an operating period is given by

$$R_m = \sum_{t \in \mathcal{T}} R_m(t). \quad (3)$$

B. Problem Formulation

In this paper, we jointly optimize user scheduling and power control to minimize the maximum access delay of all devices. Specifically, the access delay minimization problem is formulated as

$$\begin{aligned} & \min_{\mathbf{S}, \mathbf{P}, \mathcal{T}} T \\ \text{s.t.} \quad & \text{C1: } \sum_{t \in \mathcal{T}} s_m(t) = 1, \forall m \in \mathcal{U} \\ & \text{C2: } R_m \geq D_m, \forall m \in \mathcal{U} \\ & \text{C3: } s_m(t) = \{0, 1\}, \forall m \in \mathcal{U}, t \in \mathcal{T} \\ & \text{C4: } 0 \leq p_m \leq \bar{p}_m, \forall m \in \mathcal{U}. \end{aligned} \quad (4)$$

The objective function in (4) is to minimize the duration of an operating period. Access delay is usually defined as the time duration from the time a user requests a time slot to transmit till the time slot is allocated to it. Since each device has acquired a slot to transmit in the last slot of the operating period, the objective function in (4) is equivalent to minimizing the maximum access delay of all devices in the network. Besides, constraint C1 guarantees that each device is scheduled only once during an operational period. Constraints C1 and C2 together denote that the data harvested by the devices (i.e., D_m) must be transmitted to the AP in their respective scheduled slots. Constraints C3

and C4 specify the value range of the user scheduling variables (i.e., $s_m(t)$) and power control variables (i.e., p_m), where \bar{p}_m denotes the maximum transmission power of device m .

Remark 1: The problem in (4) is different from the traditional user grouping problems that have been studied in NOMA networks. First, the optimization goals are different. The existing works mainly focus on the rate-maximization or energy-minimization problems, while we aim to minimize the duration of the operating period. This problem is vital for the massive IoT scenarios, where a massive number of IoT devices collect information from surrounding environment and attempt to deliver them to the data center as soon as possible. Second, the number of user groups (NG) is given in the existing works, while it is an optimization variable in our work. This difference changes the nature of the problem. If NG is given, the problem becomes relatively easy to handle. However, it becomes more challenging when the NG is undetermined, as two correlated groups of discrete variables (i.e., the NG and the members in each group) should be jointly optimized.

III. PROBLEM ANALYSIS

In this section, we analyze the properties of the formulated problem in (4), which provide guidance for the algorithm design.

Theorem 1: The problem in (4) is NP-hard.

Proof: Assume the optimal operating period T^* and the transmission power $\mathbf{P}^* = \{p_m^* \mid m \in \mathcal{U}\}$ have been obtained in advance. Then, the problem in (4) degrades into the following user scheduling subproblem.

$$\begin{aligned} & \text{find } \mathbf{S} \\ & \text{s.t. } \text{C1: } \sum_{t \in \mathcal{T}^*} s_m(t) = 1, \forall m \in \mathcal{U} \\ & \text{C2: } R_m \geq D_m, \forall m \in \mathcal{U} \\ & \text{C3: } s_m(t) = \{0, 1\}, \forall m \in \mathcal{U}, t \in \mathcal{T}^*. \end{aligned} \quad (5)$$

In what follows, we will prove that the above problem is NP-hard.

According to (1) and constraints C1 and C2 in (5), we know that the maximum tolerable interference of device m is

$$\bar{I}_m = \frac{p_m^* G_m}{2^{\frac{D_m}{B_0}} - 1} - \sigma^2. \quad (6)$$

Note that the interference of device m comes from the devices with weaker CPG than it. Therefore, in the slot that user m is scheduled, the signal strength of the accumulative signals received by the AP (denoted as $P_{AP}(t_m)$) must obey the following inequation.

$$\begin{aligned} P_{AP}(t_m) &= \sum_{n \in \mathcal{U}} s_n(t_m) p_n^* G_n \\ &= \sum_{n \in \mathcal{U}, n > m} s_n(t_m) p_n^* G_n + \sum_{n \in \mathcal{U}, n \leq m} s_n(t_m) p_n^* G_n \\ &\leq \bar{I}_m + \sum_{n \in \mathcal{U}, n \leq m} p_n^* G_n. \end{aligned} \quad (7)$$

Base on (7), we can get that in each slot t ($t = 1, \dots, T^*$), the signal strength at the AP satisfies

$$\begin{aligned} P_{AP}(t) &= \sum_{n \in \mathcal{U}} s_n(t) p_n^* G_n \\ &\leq \max_{m \in \mathcal{U}} \left\{ \bar{I}_m + \sum_{n \in \mathcal{U}, n < m} p_n^* G_n \right\} \\ &= \Delta. \end{aligned} \quad (8)$$

Any feasible solutions of (5) satisfy the above inequations. In other words, (8) is the necessary condition of the problem in (5). As such, the simplification problem of (5) can be formulated as

$$\begin{aligned} & \text{find } \mathbf{S} \\ & \text{s.t. } \text{C1: } \sum_{t \in \mathcal{T}^*} s_m(t) = 1, \forall m \in \mathcal{U} \\ & \text{C2: } P_{AP}(t) \leq \Delta, \forall t \in \mathcal{T}^* \\ & \text{C3: } s_m(t) = \{0, 1\}, \forall m \in \mathcal{U}, t \in \mathcal{T}^*. \end{aligned} \quad (9)$$

The above problem can be described as the the following bin packing problem.

Bin Packing Problem [39]: There are M items, where the size of item m is $p_m^* G_m$. Is there a partition of the items into T^* disjoint sets, such that the total size in each set is no larger than Δ ?

The bin packing problem is a typical NP-hard problem [40]. In addition, (5) is more complicated than (9), and hence the problem in (5) is NP-hard. Since the user scheduling subproblem always exists in (4), the problem in (4) is thus NP-hard as well. To this end, we have proved Theorem 1. ■

The primal problem in (4) is intractable. To deal with it, we transform it into another form. Specifically, we have the following theorem.

Theorem 2: The problem in (4) has the same optimal solutions with the following problem.

$$\begin{aligned} & \max_{\mathbf{S}, \mathbf{P}} \sum_{t \in \mathcal{T}} \sum_{m \in \mathcal{U}} s_m(t) \\ & \text{s.t. } \text{C1: } \sum_{t \in \mathcal{T}} s_m(t) \leq 1, \forall m \in \mathcal{U} \\ & \text{C2: } R_m \geq \left(\sum_{t \in \mathcal{T}} s_m(t) \right) D_m, \forall m \in \mathcal{U} \\ & \text{C3: } s_m(t) = \{0, 1\}, \forall m \in \mathcal{U}, t \in \mathcal{T} \\ & \text{C4: } 0 \leq p_m \leq \bar{p}_m, \forall m \in \mathcal{U}, \end{aligned} \quad (10)$$

where $\mathcal{T} = \{1, 2, \dots, T^*\}$ and T^* is the optimal value of (4).

Proof: The objective of (10) is to maximize the number of devices that can be supported in a given operating period. Specifically, the operating period in (10) is set as the optimal value of (4), i.e., the minimum operating period T^* that all of the devices can be supported. As such, the maximum objective function value of the problem in (10) must be M , i.e., $\sum_{t \in \mathcal{T}} \sum_{m \in \mathcal{U}} s_m(t) = M$.

The constraint C1 in (10) indicates that $\sum_{t \in \mathcal{T}} s_m(t)$ is no larger than one for all of the device in \mathcal{U} . To make $\sum_{t \in \mathcal{T}} \sum_{m \in \mathcal{U}} s_m(t) = M$, the condition $\sum_{t \in \mathcal{T}} s_m(t) = 1$ must be satisfied for each device m , such that the constraint C1 in (10) can be rewritten as $\sum_{t \in \mathcal{T}} s_m(t) = 1, \forall m \in \mathcal{U}$. Besides, the constraint C2 in (10) can also be rewritten as $R_m \geq (\sum_{t \in \mathcal{T}} s_m(t)) D_m = D_m, \forall m \in \mathcal{U}$ accordingly. As such, the problem in (10) is recast as

$$\begin{aligned}
& \max_{\mathbf{S}, \mathbf{P}} \quad \sum_{t \in \mathcal{T}} \sum_{m \in \mathcal{U}} s_m(t) \\
& \text{s.t.} \quad \text{C1: } \sum_{t \in \mathcal{T}} s_m(t) = 1, \forall m \in \mathcal{U} \\
& \quad \quad \text{C2: } R_m \geq D_m, \forall m \in \mathcal{U} \\
& \quad \quad \text{C3: } s_m(t) = \{0, 1\}, \forall m \in \mathcal{U}, t \in \mathcal{T} \\
& \quad \quad \text{C4: } 0 \leq p_m \leq \bar{p}_m, \forall m \in \mathcal{U}. \quad (11)
\end{aligned}$$

All of the constraints in (11) are the same as those in (4), and the operating period in (11) is just equal to the optimal value of (4). Therefore, the optimal solutions of (11) (i.e., (10)) satisfy all of the constraints in (4), that is, the problems in (4) and (10) have the same optimal solutions. ■

According to Theorem 2, we can obtain the optimal user scheduling and power control policies of (4) through solving the problem in (10). But the precondition is that T^* is given in advance. In order to quickly determine the value of T^* , we design a bi-section based algorithm, which is summarized in Algorithm 1. At the step 3 in Algorithm 1, $\lceil x \rceil$ represents the smallest integer greater than or equal to x . The design philosophy of Algorithm 1 is testing whether the network can support all of the devices or not for a given value of T^* . If yes, T^* will be increased, otherwise T^* will be decreased, until the termination condition is satisfied. Specifically, the termination condition is set as $T_u - T_l = 1$. It indicates that $\sum_{t \in \mathcal{T}} \sum_{m \in \mathcal{U}} s_m(t) = M$ when $T^* = T_u$, and $\sum_{t \in \mathcal{T}} \sum_{m \in \mathcal{U}} s_m(t) < M$ when $T^* = T_l$. In this case, we can determine that T_u is the smallest operating period that can support all of the devices. In addition, the initial value of T_u is set as the number of devices, i.e., M . As such, in the worst case, T^* is equal to M , that is, each user is scheduled in one slot. The worst case must be a feasible solution of the problem in (4). Therefore, a feasible solution of (4) can always be obtained by Algorithm 1.

IV. ALGORITHM DESIGN FOR THE USER SCHEDULING AND POWER CONTROL

In this section, we aim at solving the problem in (10). Specifically, we divide it into two separated subproblems, i.e., the user scheduling subproblem (USP) and the power control subproblem (PCP). For the USP, we first remodel it as the K-CUT problem in graph theory and then propose a cost-efficient algorithm to solve it. For the PCP, we devise a low-complexity power control algorithm based on the standard interference function. Finally, an iterative resource management algorithm incorporated with the user scheduling and power control algorithms is proposed to solve the problem in (10).

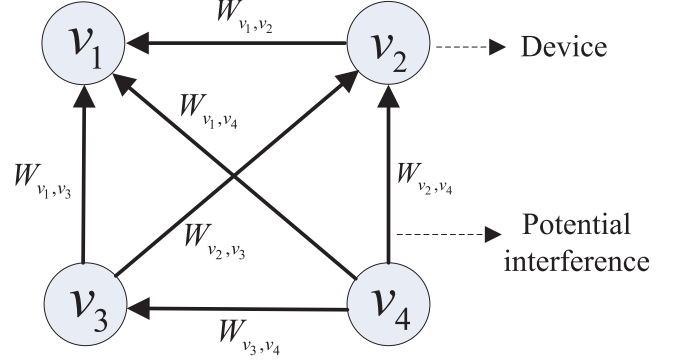


Fig. 2. A sample illustration of the interference graph.

Algorithm 1: The Overall Algorithm.

- 1: **Input:** Data volume $\{D_m\}$ and power limitation $\{\bar{p}_m\}$.
 - 2: Initialize $T_l = 1$ and $T_u = M$.
 - 3: **repeat**
 - 4: Set $T^* = \lceil \frac{T_l + T_u}{2} \rceil$.
 - 5: Solve the problem in (10) and get the control policies $\{\mathbf{S}^*, \mathbf{P}^*\}$ and objective value $\sum_{t \in \mathcal{T}} \sum_{m \in \mathcal{U}} s_m^*(t)$.
 - 6: **if** $\sum_{t \in \mathcal{T}} \sum_{m \in \mathcal{U}} s_m^*(t) < M$ **then**
 - 7: Update $T_l = T^*$.
 - 8: **else**
 - 9: Update $T_u = T^*$.
 - 10: **end if**
 - 11: **until** $T_u - T_l == 1$
 - 12: **Output:** The control policies $\{\mathbf{S}^*, \mathbf{P}^*, T^*\}$.
-

A. User Scheduling Algorithm

As proved in Theorem 1, the USP is NP-hard. As such, we cannot get the optimal solutions of (10) in polynomial time. Besides, the coupling between user scheduling and power control makes (10) more intractable. Due to these reasons, we divide (10) into the USP and the PCP and then design efficient algorithms to solve them.

In order to solve the USP efficiently, we first construct a directional interference graph to model the interference among devices. Specifically, we take the case with four devices as an example to illustrate the interference graph $G(V, E)$. As depicted in Fig. 2, the vertices V in the graph represent the devices, and the directed edges E demonstrate the interference among devices. For each edge $(u, v) \in E$, we define a weight $W_{u,v}$ to represent the mutual interference between the two users, which is specified as

$$W_{u,v} = \begin{cases} \bar{p}_u G_u, & \text{if } G_u < G_v \\ \bar{p}_v G_v, & \text{if } G_u > G_v \end{cases}. \quad (12)$$

Noted that in an unidirectional interference graph, the mutual interference is usually calculated by $\bar{p}_u G_u + \bar{p}_v G_v$. Thus, the directional interference is considered in our model. Intra-cell interference is the main factor that limits the performance of NOMA networks, while user scheduling is an effective measure

of interference coordination. Appropriate scheduling policy can eliminate the strong interference and thereby improve the number of connections. In light of these, we remodel the USP as the interference coordination problem, which is described as the following K-CUT problem.

Definition 1: (K-CUT): Given a graph $G(V, E)$, delete partial edges and split $G(V, E)$ into K disconnected sub-graphs $G_1(Q_1, E_1), G_2(Q_2, E_2), \dots, G_K(Q_K, E_K)$, where $\cup_{k=1}^K Q_k = V$. Then, the K sub-graphs are called as a K-CUT of the graph $G(V, E)$.

Definition 2: (K-CUT Problem): Given an interference graph $G(V, E)$, the K-CUT problem is to find the K-CUT from all possible solutions, such that the removed interference $O = \sum_{i=1}^K \sum_{j=i+1}^K (\sum_{u \in Q_i} \sum_{v \in Q_j} W_{u,v})$ is maximized.

In the K-CUT problem, O (i.e., the sum-interference among different clusters) is maximized, that is, the remainder interference is minimized. Thus, the K-CUT problem well models the interference coordination problem in NOMA networks. Based on the methods in graph theory [39], we can design a cost-efficient algorithm to solve the formulated interference coordination problem.

To denote the sum-interference between vertex u and cluster Q_k , we define I_{u, Q_k} as

$$I_{u, Q_k} = \sum_{v \in Q_k, v \neq u} W_{u,v}. \quad (13)$$

The sum-interference between two different clusters Q_i and Q_j is defined as

$$I_{Q_i, Q_j} = \sum_{u \in Q_i} \sum_{v \in Q_j} W_{u,v} = \sum_{u \in Q_i} I_{u, Q_j} = \sum_{v \in Q_j} I_{v, Q_i}. \quad (14)$$

Furthermore, the sum-interference in the same cluster Q_i is defined as

$$I_{Q_i, Q_i} = \sum_{u=1}^{|\mathcal{Q}_i|} \sum_{v=u+1}^{|\mathcal{Q}_i|} W_{u,v} = \frac{1}{2} \sum_{u \in Q_i} I_{u, Q_i}. \quad (15)$$

The user scheduling algorithm, referred to as USA, is summarized in Algorithm 2 and detailed as follows. There are two main stages in the USA. In the first stage (i.e., steps 2–9), K clusters are constructed by a heuristic method. In the second stage (i.e., steps 10–17), the vertexes are transferred among the K clusters so as to reduce the total interference. By this way, the proposed algorithm can well coordinate the intra-cell interference. Denote O^* as the optimal value of the K-CUT problem, that is, O^* is equal to the maximum value of $\sum_{i=1}^K \sum_{j=i+1}^K (\sum_{u \in Q_i} \sum_{v \in Q_j} W_{u,v})$ among all possible K-CUT solutions. Then, we can get the following conclusion on the worst-case performance of the proposed USA.

Theorem 3: The ratio of O to O^* is no smaller than $1 - \frac{1}{K}$, i.e., $\frac{O}{O^*} \geq 1 - \frac{1}{K}$.

Proof: According to the definitions of O and I_{Q_i, Q_j} , it can be obtained that

$$O = \sum_{i=1}^K \sum_{j=i+1}^K \left(\sum_{u \in Q_i} \sum_{v \in Q_j} W_{u,v} \right) = \sum_{i=1}^K \sum_{j=i+1}^K I_{Q_i, Q_j}. \quad (16)$$

Besides, the termination condition of Algorithm 2 indicates that the following inequation must hold at the end of the algorithm.

$$I_{u, Q_i} \leq I_{u, Q_j}, \forall u \in Q_i, i \neq j. \quad (17)$$

Summing up both sides of (17) over all $u \in Q_i$ yields

$$\sum_{u \in Q_i} I_{u, Q_i} \leq \sum_{u \in Q_i} I_{u, Q_j}. \quad (18)$$

According to (14) and (15), we can rewrite (18) as

$$2I_{Q_i, Q_i} \leq I_{Q_i, Q_j}. \quad (19)$$

Summing up both sides of (19) over all $i, j = 1, \dots, K, i \neq j$ yields

$$2 \sum_{i=1}^K \sum_{j=1, i \neq j}^K I_{Q_i, Q_i} \leq \sum_{i=1}^K \sum_{j=1, i \neq j}^K I_{Q_i, Q_j}. \quad (20)$$

According to (16), the above inequation can be transformed into

$$2(K-1) \sum_{i=1}^K I_{Q_i, Q_i} \leq 2O. \quad (21)$$

Rearranging (21), we can get

$$\sum_{i=1}^K I_{Q_i, Q_i} \leq \frac{O}{K-1}. \quad (22)$$

Since the optimal value of the K-CUT problem (i.e., O^*) must be smaller than or equal to the total interference in the interference graph, it thus holds that

$$O^* \leq \sum_{i=1}^K I_{Q_i, Q_i} + O \leq \frac{O}{K-1} + O. \quad (23)$$

Rearranging (23), we can get

$$\frac{O}{O^*} \geq 1 - \frac{1}{K}. \quad (24)$$

To this end, we have proofed Theorem 3. \blacksquare

The K-CUT problem aims to maximize the removed interference, which is equivalent to minimizing the remaining interference. Theorem 3 indicates that the USA can achieve at least $(1 - \frac{1}{K}) O^*$. Thus, the worst-case performance of the USA is presented by Theorem 3. For instance, if $K = 2$, the USA can achieve at least $\frac{O^*}{2}$. When K is large, a better performance can be obtained. Denote U as the remaining interference obtained by the USA, i.e., $U = \sum_i I_{Q_i, Q_i}$. Define U^* as the minimum value of $\sum_i I_{Q_i, Q_i}$ and $U^* = \lambda O^*$, where λ is a constant for a given interference graph. Then, based on the conclusion of Theorem 3, we can get that $U \leq (1 + \frac{1}{\lambda K}) U^*$.

B. Power Control Algorithm

In the K-CUT problem, each cluster corresponds to a time slot. Given the clusters, we should further optimize the transmission power of the devices to satisfy their rate requirements. Since the devices in different clusters are independent, we can solve

Algorithm 2: User Scheduling Algorithm (USA).

-
- 1: **Input:** User set \mathcal{B} and cluster number K .
 - 2: Construct the interference graph $G(V, E)$.
 - 3: Choose K vertexes in V randomly as K initial clusters, denoted by $\mathcal{Q}_1, \mathcal{Q}_2, \dots, \mathcal{Q}_K$ respectively.
 - 4: Set $V \leftarrow V \setminus \mathcal{Q}_1 \setminus \mathcal{Q}_2 \setminus \dots \setminus \mathcal{Q}_K$.
 - 5: **repeat**
 - 6: Choose one vertex in V randomly, denoted by v^* .
 - 7: Choose the cluster k^* from the K clusters according to $k^* = \arg \min_{k=1, \dots, K} I_{v^*, \mathcal{Q}_k}$.
 - 8: Set $\mathcal{Q}_{k^*} \leftarrow \mathcal{Q}_{k^*} \cup \{v^*\}$ and $V \leftarrow V \setminus \{v^*\}$.
 - 9: **until** $V == \emptyset$
 - 10: **while** 1 **do**
 - 11: Set $R = 1$ if there is a vertex $v^* \in \mathcal{Q}_i$ and a cluster \mathcal{Q}_j where $i \neq j$ and $I_{v^*, \mathcal{Q}_i} > I_{v^*, \mathcal{Q}_j}$, otherwise set $R = 0$.
 - 12: **if** $R == 1$ **then**
 - 13: Move v^* from \mathcal{Q}_i to \mathcal{Q}_{k^*} , where $k^* = \arg \min_{k=1, \dots, K} I_{v^*, \mathcal{Q}_k}$.
 - 14: **else**
 - 15: Break.
 - 16: **end if**
 - 17: **end while**
 - 18: **Output:** The K clusters $\mathcal{Q}_1, \mathcal{Q}_2, \dots, \mathcal{Q}_K$.
-

the PCP for each cluster separately. Specifically, the achievable data rate of device m in \mathcal{Q}_k can be rewritten as

$$R_m = B_0 \log_2 \left(1 + \frac{p_m G_m}{\sum_{n \in \mathcal{Q}_k, n > m} p_n G_n + \sigma^2} \right). \quad (25)$$

The PCP in \mathcal{Q}_k can be formulated as

$$\begin{aligned} \text{find } & \mathbf{P}_k = \{p_m \mid m \in \mathcal{Q}_k\} \\ \text{s.t. } & \text{C1: } R_m \geq D_m, \forall m \in \mathcal{Q}_k \\ & \text{C2: } 0 \leq p_m \leq \bar{p}_m, \forall m \in \mathcal{Q}_k. \end{aligned} \quad (26)$$

Noted that the above problem may be infeasible due to the constrains C1 and C2. Thus, it is very hard to tackle (26) directly. To make the PCP always feasible, we formulate the following problem.

$$\begin{aligned} \min_{\mathbf{P}_k} \max_{m \in \mathcal{Q}_k} & \frac{p_m}{\bar{p}_m} \\ \text{s.t. } & \text{C1: } R_m \geq D_m, \forall m \in \mathcal{Q}_k \\ & \text{C2: } p_m \geq 0, \forall m \in \mathcal{Q}_k. \end{aligned} \quad (27)$$

The relationship between (26) and (27) is demonstrated as the following theorem.

Theorem 4: If the optimal value of (27) is no larger than one, the problem in (26) is feasible, and the optimal solutions of (27) are also the feasible solutions of (26), otherwise the problem in (26) is infeasible.

Proof: First, we prove that the problem in (27) is always feasible. Combining (25) and C1 in (27), we can get the minimum

transmission power of device m that satisfies its minimum data rate requirement under the condition that the transmission power of other devices is fixed. This minimum value is specified as

$$X_m(\mathbf{P}_k) = \frac{\left(2^{\frac{D_m}{B_0}} - 1\right) \left(\sum_{n \in \mathcal{Q}_k, n > m} p_n G_n + \sigma^2\right)}{G_m}. \quad (28)$$

Define $\mathbf{X}(\mathbf{P}_k) = \{X_m(\mathbf{P}_k) \mid m \in \mathcal{Q}_k\}$. We can easily proof that $\mathbf{X}(\mathbf{P}_k)$ has the following three properties:

- 1) Positivity: For any \mathbf{P}_k , $\mathbf{X}(\mathbf{P}_k) > 0$ holds.
- 2) Monotonicity: If $\mathbf{P}_k \geq \mathbf{P}'_k$, then $\mathbf{X}(\mathbf{P}_k) \geq \mathbf{X}(\mathbf{P}'_k)$.
- 3) Scalability: For any $\alpha > 1$, $\alpha \mathbf{X}(\mathbf{P}_k) > \mathbf{X}(\alpha \mathbf{P}_k)$ holds.

It has been proved in [41] that if a vector satisfies the positivity, monotonicity, and scalability, this vector is a standard interference function, thereby $\mathbf{X}(\mathbf{P}_k)$ is a standard interference function. If an interference function is standard, the global optimal solution must exist and can be obtained in an iterative manner [41]. Since there is no upper limit for \mathbf{P}_k in (27), the constraint C1 in (27) can be satisfied for any D_m , that is, the problem in (27) is always feasible.

If the optimal value of (27) is no larger than one, the condition of $0 \leq p_m \leq \bar{p}_m$ holds for each $m \in \mathcal{Q}_k$. The corresponding solutions of (27) are also the feasible solutions of (26). On the contrary, if the optimal value of (27) is larger than one, there is at least one device which cannot satisfy its rate requirement under the power constraint $0 \leq p_m \leq \bar{p}_m$. In this case, the problem in (26) is infeasible. ■

According to Theorem 4, we can obtain the feasible solutions of (26) or confirm that it is infeasible by solving the problem in (27). Furthermore, based on the standard interference function, we propose an iterative power control algorithm to solve the problem in (27), which corresponds to the steps 3–8 in Algorithm 3. According to the obtained results, we further deal with the problem in (26). To make (26) feasible, we adopt the admission control scheme based on Theorem 4. Specifically, the device with the maximum power ratio (i.e., $m^* = \arg \max_{m \in \mathcal{Q}_k} \frac{p_m}{\bar{p}_m}$) is removed one by one until the objective function of (27) is no larger than 1 (i.e., $f^i = \max_{m \in \mathcal{Q}_k} \frac{p_m}{\bar{p}_m} \leq 1$). These operations correspond to the steps 9–13 in Algorithm 3. When $f^i \leq 1$, the problem in (26) becomes feasible, that is, the remaining devices can coexist in the same slot. Finally, we can get the user scheduling and power control policies in slot t according to (29) and (30). In addition, the removed devices will be assigned to other clusters, which will be discussed in the following subsection.

C. Iterative Algorithm and General Procedure

The USA and PCA focus on the user scheduling subproblem and power control subproblem, respectively. In this section, we propose an iterative algorithm to solve the joint optimization problem in (10). The proposed resource management algorithm, referred to as IRMA, is given in Algorithm 4. Specifically, the IRMA optimizes user scheduling and power control in an iterative manner. In each iteration, the users are grouped into K clusters by using the USA (step 5). Then, the cluster with the most members is selected (step 6), and the PCA is adopted to optimize the user's transmission power (step 7). By this way, a

Algorithm 3: Power Control Algorithm (PCA).

-
- 1: **Input:** User cluster \mathcal{Q}_k .
 - 2: **while** 1 **do**
 - 3: Set $i = 1$, $f^i = 0$, and $\mathbf{P}_k = \{p_m = \bar{p}_m \mid m \in \mathcal{Q}_k\}$.
 - 4: **repeat**
 - 5: Set $i = i + 1$.
 - 6: Update the transmission power of each user $m \in \mathcal{Q}_k$ according to $p_m = X_m(\mathbf{P}_k)$.
 - 7: Calculate $f^i = \max_{m \in \mathcal{Q}_k} \frac{p_m}{\bar{p}_m}$.
 - 8: **until** $|f^i - f^{i-1}| \leq \varepsilon$
 - 9: **if** $f^i \leq 1$ **then**
 - 10: Break;
 - 11: **else**
 - 12: Find the user m^* according to $m^* = \arg \max_{m \in \mathcal{Q}_k} \frac{p_m}{\bar{p}_m}$ and remove it from \mathcal{Q}_k , i.e., $\mathcal{Q}_k \leftarrow \mathcal{Q}_k \setminus \{m^*\}$;
 - 13: **end if**
 - 14: **end while**
 - 15: Get the control policy $\{\mathbf{S}^*(t), \mathbf{P}_k^*\}$ according to

$$s_m^*(t) = \begin{cases} 1, & \text{if } m \in \mathcal{Q}_k, \forall m \in \mathcal{U}. \\ 0, & \text{otherwise} \end{cases} \quad (29)$$

$$p_m^* = \begin{cases} p_m, & \text{if } m \in \mathcal{Q}_k, \forall m \in \mathcal{U}. \\ 0, & \text{otherwise} \end{cases} \quad (30)$$

-
- 16: **Output:** The control policy $\{\mathbf{S}^*(t), \mathbf{P}_k^*\}$.
-

feasible user scheduling and power control policy in slot t is obtained. In the next iteration, t is increased by one, i.e., $t = t + 1$ (step 3). Based on the previous control policy, the remaining users are regrouped into $K - 1$ clusters by the USA, and the transmission power of the users in the cluster with most members is optimized by the PCA. The aforementioned operations are repeated until $t = T^*$. As such, we can get the user scheduling and power control policies in all slots after T^* iterations. Since the USA and the PCA are iteratively implemented, a better solution of the problem in (10) can be obtained at the end of the IRMA in comparison with the independent operation of the USA and the PCA.

The general procedure of our proposed scheme for solving the problem in (4) is illustrated in Fig. 3. More detailedly, Algorithm 1 chooses the operating period T^* by a bisection method and puts it into Algorithm 4. By calling Algorithm 2 and Algorithm 3 in an iterative manner, Algorithm 4 outputs the control policy $\{\mathbf{S}^*, \mathbf{P}^*\}$ for the given operating period T^* . Then, Algorithm 1 tests whether all of the devices can be supported under the control policy $\{\mathbf{S}^*, \mathbf{P}^*\}$ or not. If yes, the final control policy $\{\mathbf{S}^*, \mathbf{P}^*, T^*\}$ is obtained, otherwise the aforementioned operations are repeated until the termination condition is satisfied.

The problem considered in the paper is NP-hard, which is very hard to tackle. To solve it effectively, a tradeoff between performance and complexity should be made. In order to get a good solution, multiple iterative operations are implemented in our

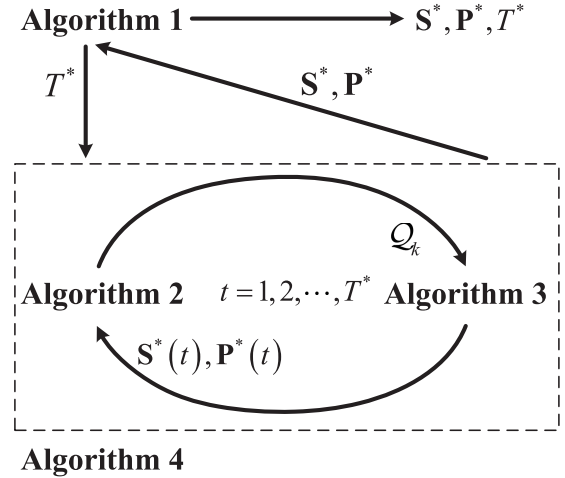


Fig. 3. General procedure of the proposed scheme.

Algorithm 4: Iterative Resource Management Algorithm (IRMA).

-
- 1: **Input:** Operating period T^* .
 - 2: Set $\mathcal{B} = \mathcal{U}$ and $\mathcal{C} = \emptyset$.
 - 3: **for** $t = 1 : 1 : T^*$ **do**
 - 4: Set $\mathcal{B} \leftarrow \mathcal{B} \setminus \mathcal{C}$ and $K = T^* - t + 1$;
 - 5: Put \mathcal{B} and K into Algorithm 2 and get $\mathcal{Q}_1, \mathcal{Q}_2, \dots, \mathcal{Q}_K$.
 - 6: Choose the cluster \mathcal{Q}_k according to $k = \arg \max_{i=1, \dots, K} |\mathcal{Q}_i|$.
 - 7: Put \mathcal{Q}_k into Algorithm 3 and get $\{\mathbf{S}^*(t), \mathbf{P}_k^*\}$.
 - 8: Get the user set $\mathcal{C} = \{m \mid m \in \mathcal{U}, s_m^*(t) = 1\}$.
 - 9: **end for**
 - 10: **Output:** The control policy $\{\mathbf{S}^*, \mathbf{P}^*\}$.
-

proposed algorithms, which will consume some computing time. As a consequence, there are some constraints for the application scenario of our algorithms. In general, our algorithms can be applied into two types of scenarios. The first one is that the position of devices and the surrounding environment are fairly static, e.g., some sensor networks. In this scenario, the transmission demand and the channel condition of the devices remain about the same or change very slowly. For this type of application scenario, we can implement the proposed algorithms one time at the beginning of the system or periodically during runtime. The other scenario is that the transmission demand and the channel condition of the devices changes relatively quickly, however the traffic is delay-insensitive, e.g., file transfer of some slow-moving devices. For this scenario, our algorithms can also be applied.

V. SIMULATION RESULTS

In this section, we present plenty of simulation results to investigate the performance of our proposed algorithms. Specifically, we conduct simulations of the NOMA network with cell radius as 1000 m, wherein the IoT devices are randomly

and uniformly deployed in the coverage area of the AP. The channel bandwidth is set as 180 KHz. The noise power is set as 1.8×10^{-14} W, that is, the noise power spectrum density (NPSD) is -160 dBm/Hz. Besides, the large-scale channel fading including path loss and shadowing fading is considered in the simulations. Detailedly, the path loss is calculated by $128.1 + 37.6 \log_{10}(d[\text{km}])$, and the shadowing fading is randomly generated according to a log-normal distribution with mean as zero and variance as eight. Other parameters are specified in each simulation.

To demonstrate the advantages of our scheme, we compare it with other six schemes, namely the US-Only, PC-Only, Random, MTS [11], DPC [21], and OMA. The detailed introduction for these four schemes are given in the following.

- US-Only: This scheme adopts our general optimization procedure (i.e., Algorithm 1 and Algorithm 4) and the user scheduling algorithm (i.e., Algorithm 2) but without power control. The transmission power of the devices are set as their maximum values, i.e., $p_m = \bar{p}_m, \forall m \in \mathcal{U}$.
- PC-Only: This scheme employs our proposed power control algorithm (i.e., Algorithm 3) and the general optimization procedure. The difference is that the devices are randomly scheduled by the PC-Only scheme among the slots. Specifically, the devices are randomly divided into T groups, each of which corresponds to a slot.
- Random: This scheme also adopts our general optimization procedure but without optimizing the transmission power and the scheduling scheme of the devices. In particular, the devices are randomly assigned to T slots with their maximum transmission power.
- MTS: The MTS proposed in [11] is a user scheduling algorithm, which schedules users into n given slots to maximize the throughput of the network. To make it applicable for our problem, we embed it into Algorithm 1 and replace the step 5. This algorithm can be used to evaluate the performance of our proposed user scheduling algorithm (i.e., Algorithm 2).
- DPC: The DPC proposed in [21] is a distributed power control algorithm, which aims to minimize the total power consumption through optimize the transmission power. Similar to the PC-Only, the devices are also randomly scheduled among slots. By comparing with the DPC, we can evaluate the performance of our proposed power control algorithm (i.e., Algorithm 3).
- OMA: Different from the former schemes, the OMA scheme adopts the orthogonal multiple access technique. With this scheme, the minimum operating period is identical to the number of devices. The OMA scheme is utilized as a benchmark to illustrate the performance gain of the NOMA network and our proposed algorithms.

A. Convergence Performance

In this subsection, we investigate the convergence property of our proposed Algorithms 1, 2, and 3.

Fig. 4 shows the convergence evolution of Algorithm 1. The simulation results corresponding to each line (i.e., each M)

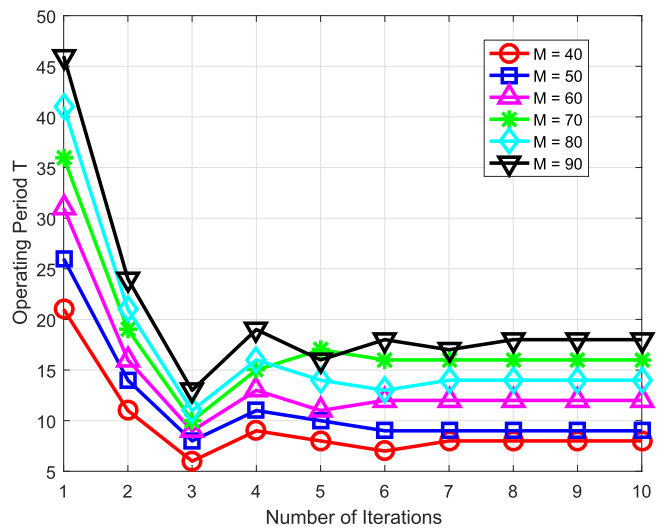


Fig. 4. Convergence evolution of Algorithm 1 under different number of devices ($\bar{p}_m = 200$ mW, $D_m = 500$ Kbits, $\forall m \in \mathcal{U}$).

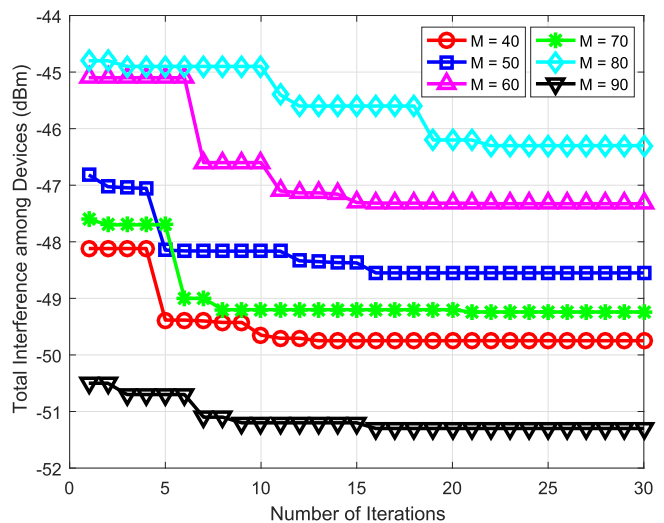


Fig. 5. Convergence evolution of Algorithm 2 under different number of devices ($\bar{p}_m = 200$ mW, $D_m = 500$ Kbits, $\forall m \in \mathcal{U}$).

are obtained by a random realization. From this figure, it can be observed that the operating period under each number of devices can converge to a stable value rapidly, and the number of iterations is usually smaller than ten. This result confirms the fast convergence property of Algorithm 1. Furthermore, we can find that the operating period when $M = 70$ is larger than that when $M = 80$. This result indicates that the minimum operating period is not only dependent on the number of devices but also affected by the channel conditions of the devices. Therefore, to fully exploit the advantages of NOMA and thereby minimize the access delay, the specific channel conditions of the devices should be taken into account when designing the power control and user scheduling schemes.

Fig. 5 plots the convergence evolution of Algorithm 2, which shows the total interference versus the number of iterations. As described in subSection V-A, the user scheduling subproblem is remodeled as a K-CUT problem with the objective to minimize

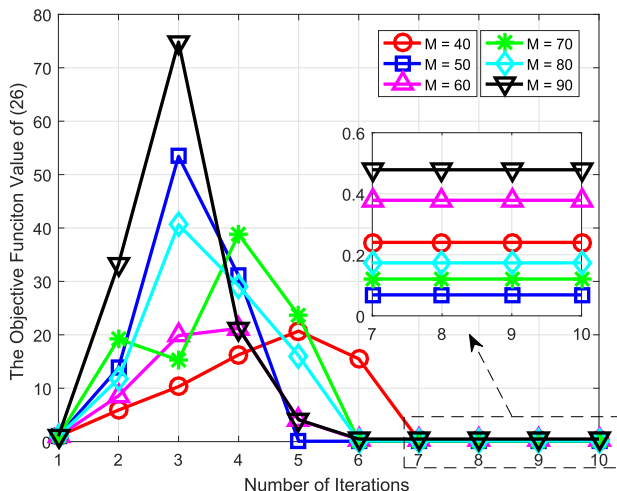


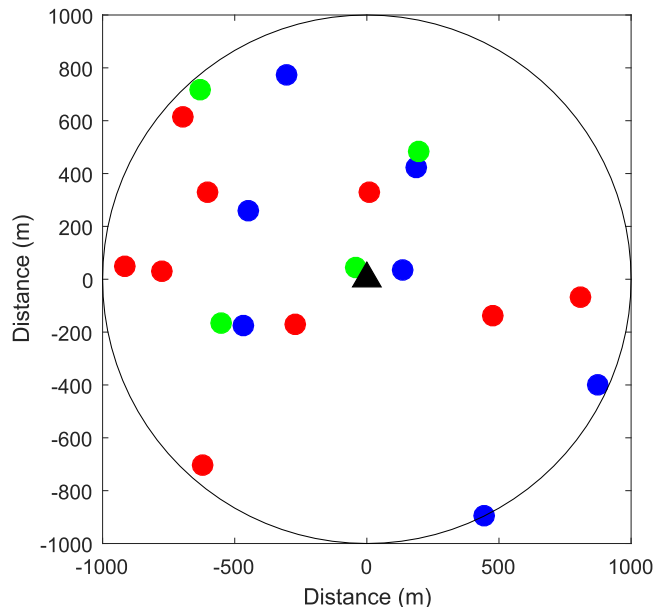
Fig. 6. Convergence evolution of Algorithm 3 under different number of devices ($\bar{p}_m = 200$ mW, $D_m = 500$ Kbits, $\forall m \in \mathcal{U}$).

the total interference among devices. Motivated by the K-CUT problem, we propose the Algorithm 2. A key operation in Algorithm 2 is step 11, which removes a device from one cluster to another if the total interference can be reduced. With the execution of the algorithm, the total interference will be reduced gradually until reaching a certain value. This variation trend is verified by Fig. 5. Besides, this figure shows that the number of iterations under different number of devices is usually smaller than 30. Thus, Algorithm 4 also has good convergence, which is helpful for its application in practical systems.

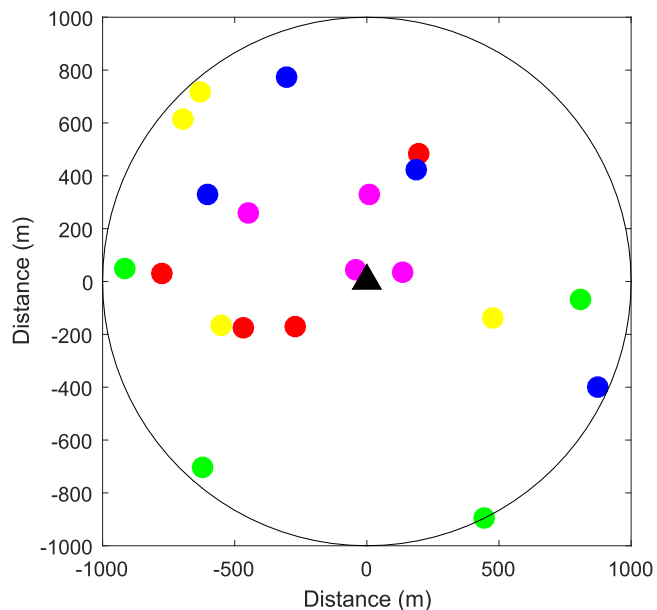
Fig. 6 demonstrates the convergence evolution of Algorithm 3, where the y-axis denotes the objective function value (OFV) of (27). To solve the problem in (27), we propose an iterative power control algorithm based on the standard interference function. As shown in Fig. 6, the OFV of (27) converges to a fixed point only after several iterations. This result is consistent with the theory of interference function, which shows that if an interference function is standard, the iterative algorithm can converge geometrically fast to the global optimal value of the problem with any initial points. Additionally, this figure shows that the OFV increases first and then decreases during the iterative procedure. The main reason for this phenomenon is due to the directional interference in NOMA networks. More specifically, affected by the decoding order of SIC, the near devices (with large CPG) undergo large interference while the far devices (with small CPG) undergo small interference. In the initial stage, the near devices will improve their transmission power to satisfy the rate requirements, while the far devices will decrease their transmission power to save energy. After several iterations, the interference received by the near devices is reduced due to the decrement of the transmission power of the far devices. Consequently, the transmission power of the near devices will also decrease with iterations until reaching the stable state.

B. Snapshots of Scheduling Results

To give some insight into our proposed user scheduling algorithm, we plot the scheduling results of our algorithm and the



(a) Our Algorithm ($T=3$)



(b) Random Scheduling Algorithm ($T=5$)

Fig. 7. Scheduling results under different algorithms, where each color represents a slot ($M = 20$, $\bar{p}_m = 200$ mW, $D_m = 300$ kbit, $\forall m \in \mathcal{U}$).

PC-Only scheme (i.e., a random scheduling algorithm). The simulation results are obtained by a random realization and shown in Fig. 7, where the dots and the triangle denote the devices and the AP respectively, and different colors of the dots represent different time slots. As shown in this figure, the operating period T with our algorithm is three, while T is five with the random scheduling algorithm. Thus, our algorithm greatly outperforms the random scheduling algorithm in terms of T . From Fig. 7(a), we can find that there are both near devices and far devices in each slot with our algorithm. By this way, the user diversity in the power domain is fully exploited, and hence more devices can

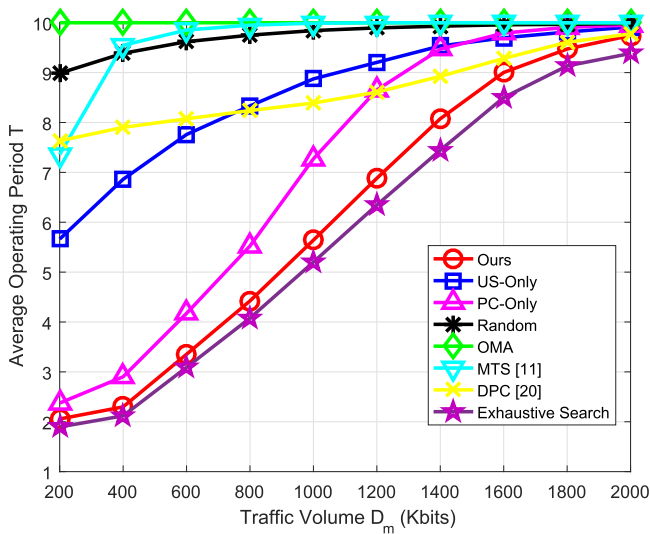


Fig. 8. Average operating period under different traffic volume D_m ($\bar{p}_m = 200$ mW, $M = 10$).

be supported in the same slot. As a consequence, our algorithm can satisfy the traffic demand of all devices in a short operating period. By contrast, with the random user scheduling algorithm, the distribution of the devices in different slots is very uneven, as depicted in Fig. 7(b). For instance, all green dots are distributed in the cell-edge area, while the pink dots gather around the AP. Due to the uneven distribution, each slot can only accommodate a small number of devices, thereby leading to a long operating period. The aforementioned reasons account for why our algorithm outperforms the random scheduling algorithm.

C. Performance Comparison

In this subsection, we evaluate the performance of our scheme by comparing it with the US-Only, PC-Only, Random, OMA, MTS [11], DPC [21], and exhaustive search method under different system parameters. The resulting values are obtained by averaging over 2000 random simulation runs.

To evaluate the gap between our solution and the optimal solution, we first compare our scheme with the exhaustive search method in a small-scale scenario. Specifically, we set the number of devices as 10 and change the traffic volume from 200 kbits to 2 Mbits. It is noted that even in this small-scale scenario, the complexity of the exhaustive search method is still very high. This is because we should test the operating period T from 1 to 10, and furthermore for each given T , we should check the feasibility for every possible user scheduling scheme. As shown in Fig. 8, the gap between our scheme and the exhaustive search method is small. The data shows that our scheme only increases about 8% operating period with respect to the optimal solution. For instance, if the optimal operating period is 25 slots, our scheme only consumes about two more slots. However, the computational complexity is reduced significantly, which makes our scheme applicable in practical networks. In addition, it can be observed that even in this small-scale scenario, our scheme can still achieve large performance gain in comparison with

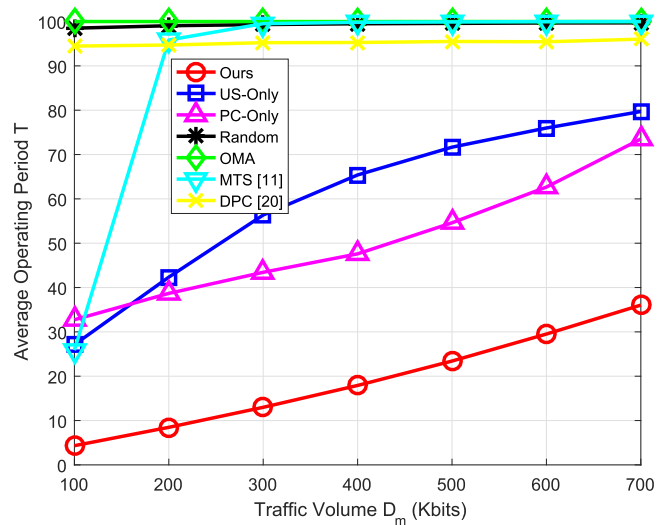


Fig. 9. Average operating period under different traffic volume D_m ($\bar{p}_m = 200$ mW, $M = 100$).

other schemes. These results verify the effectiveness of our proposed algorithms.

Fig. 9 shows the effect of different traffic volume D_m on the average operating period (AOP). First, we can find that our scheme can greatly reduce the AOP with respect to other schemes. Besides, the US-Only and the PC-Only outperform the MTS and the DPC. This result verifies that both our proposed power control algorithm and user scheduling algorithm are very effective. Since the MTS and the DPC respectively focus on throughput maximization and power minimization, they exhibit bad performance on the AOP. This reflects that to minimize the access delay of devices, new user scheduling and power control algorithms should be designed. Furthermore, this figure shows that the NOMA networks with random scheme almost have the same performance with the OMA networks in terms of the AOP. However, if the transmission power or the scheduling scheme of the devices is optimized, the NOMA networks can exhibit its advantages on the access delay (or device connections). Moreover, the simulation results also tell us that both the power control and the user scheduling are efficiency control policies, which are capable of enhancing the performance of NOMA networks significantly. Therefore, to take the advantages of NOMA, it is necessary to optimize the transmission power of the devices incorporated with appropriate user scheduling scheme.

Fig. 10 plots the AOP versus the number of devices. As can be seen, the AOP with all schemes almost linearly increases with the number of devices. When there exists abundant devices in the network, the access delay with the OMA scheme becomes very large. This result indicates that the traditional OMA techniques are not suitable for the network with massive IoT devices. Besides, if the resources (slots and power) are not properly allocated among the devices, the AOP of the NOMA network is also very large. As can be seen, the performance of the MTS, the DPC, and the Random is almost the same as the OMA. By contrast, the NOMA networks with our scheme can keep the AOP at a relatively low level. For instance, the AOP is about 25

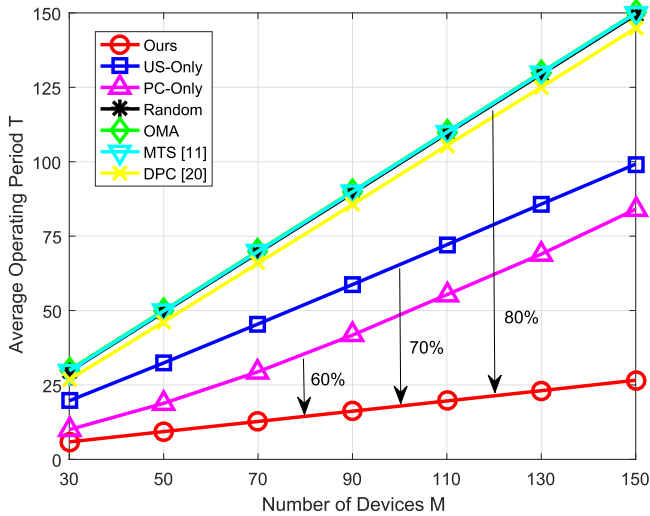


Fig. 10. Average operating period under different number of devices M ($\bar{p}_m = 200$ mW, $D_m = 400$ Kbits, $\forall m \in \mathcal{U}$).

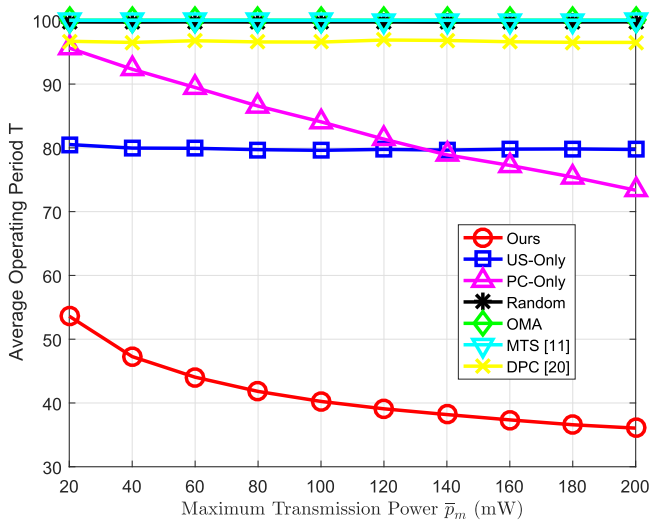


Fig. 11. Average operating period under different maximum transmission power constraints \bar{p}_m ($M = 100$, $D_m = 700$ Kbits, $\forall m \in \mathcal{U}$).

when there exists 150 devices in the NOMA network, that is, each slot can support 6 devices. This result demonstrates that the NOMA networks with appropriate optimization algorithms can be applied into the 5G application scenarios of low latency and massive connections. In addition, from Fig. 9 and Fig. 10, we can see that the power control policy can achieve more performance gain in comparison with the user scheduling policy when the traffic demand is large, while the situation is reverse when the traffic demand is small. Thus, if the operating complexity of the network is limited, a decision on how to choose control policy should be made according to the traffic demand.

Fig. 11 shows the effect of the maximum transmission power \bar{p}_m on the AOP. It can be observed from this figure that the AOP with the OMA, US-Only, MTS, DPC, and Random schemes is almost unchanged with the increment of \bar{p}_m . However, \bar{p}_m has a great effect on our scheme and the PC-Only scheme. The reason

is specified as follows. To make the problem in (4) feasible, we assume that the traffic demand can be satisfied if each device is scheduled in an individual slot with its maximum transmission power. Hence, different values of \bar{p}_m have no effect on the AOP with the OMA scheme. As mentioned before, the MTS, DPC, and Random schemes have the similar performance with the OMA scheme, thus \bar{p}_m also has no effect on it. Besides, the transmission power of the devices with the US-Only scheme is set as the maximum value. As such, the SINR of the devices is almost unchanged when the transmission power of the devices is improved proportionally. Accordingly, the AOP with the US-Only scheme also keeps unchanged under different values of \bar{p}_m . On the contrary, our scheme and the PC-Only scheme set different transmission power for each device. It is noted that the performance gain derived by the power control policy comes from the user diversity in the power domain. If \bar{p}_m is large, the transmission power of the devices can be set as more different levels, and hence more devices can be supported in the same slot by NOMA technique. Therefore, the effect of \bar{p}_m on the AOP with our scheme and the PC-Only scheme is more obvious.

VI. CONCLUSION

This paper has investigated the access delay minimization problem (ADMP) for the uplink NOMA networks with massive connections by jointly considering user scheduling and power control. We have formulated the ADMP as a mixed-integer and non-convex programming problem, which has been proved to be NP-hard. To tackle this hard problem, we have proposed a low-complexity algorithm based on the K-CUT method and standard interference function. Specifically, the proposed algorithm solves the user scheduling subproblem and the power control subproblem in an iterative manner. Finally, we have conducted abundant simulations to evaluate the performance of our algorithm. The simulation results have verified the good performance of our algorithm in terms of convergence and access delay.

REFERENCES

- [1] "Internet of things," The China Academy of Information and Communications Technology (CAICT), Beijing, China, White Paper, Dec. 2016.
- [2] "5G vision and requirements," IMT-2020 5G Promotion Group, Beijing, China, White Paper, May 2014.
- [3] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. IEEE VTC'13-Spring*, Dresden, Germany, Jun. 2013, pp. 1–5.
- [4] L. Dai, B. Wang, Y. Yuan, S. Han, C. I, and Z. Wang, "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.
- [5] "5G wireless technology architecture," IMT-2020 5G Promotion Group, Beijing, China, White Paper, May 2015.
- [6] Z. Ding *et al.*, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.
- [7] M. Shirvanimoghaddam, M. Dohler, and S. J. Johnson, "Massive non-orthogonal multiple access for cellular IoT: Potentials and limitations," *IEEE Commun. Mag.*, vol. 55, no. 9, pp. 55–61, Sep. 2017.
- [8] Z. Yang, Z. Ding, P. Fan, and N. Al-Dhahir, "A general power allocation scheme to guarantee quality of service in downlink and uplink NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7244–7257, Nov. 2016.

- [9] J. Choi, "On the power allocation for MIMO-NOMA systems with layered transmissions," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3226–3237, May 2016.
- [10] X. Wang and L. Cai, "Proportional fair scheduling in hierarchical modulation aided wireless networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 4, pp. 1584–1593, Apr. 2013.
- [11] M. Mollanouri and M. Ghaderi, "Uplink scheduling in wireless networks with successive interference cancellation," *IEEE Trans. Mobile Comput.*, vol. 13, no. 5, pp. 1132–1144, May 2014.
- [12] 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.
- [13] J. Zhao, Y. Liu, K. K. Chai, A. Nallanathan, Y. Chen, and Z. Han, "Spectrum allocation and power control for non-orthogonal multiple access in HetNets," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 5825–5837, Sep. 2017.
- [14] D. Zhai and J. Du, "Spectrum efficient resource management for multi-carrier based NOMA networks: A graph-based method," *IEEE Wireless Commun. Lett.*, vol. 7, no. 3, pp. 388–391, Jun. 2018.
- [15] L. Lei, D. Yuan, C. K. Ho, and S. Sun, "Power and channel allocation for non-orthogonal multiple access in 5G systems: Tractability and computation," *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 8580–8594, Dec. 2016.
- [16] W. Bao, H. Chen, Y. Li, and B. Vucetic, "Joint rate control and power allocation for non-orthogonal multiple access systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2798–2811, Dec. 2017.
- [17] M. S. Ali, H. Tabassum, and E. Hossain, "Dynamic user clustering and power allocation for uplink and downlink non-orthogonal multiple access (NOMA) systems," *IEEE Access*, vol. 4, pp. 6325–6343, Aug. 2016.
- [18] D. Wang, P. Ren, Q. Du, L. Sun, and Y. Wang, "Security provisioning for MISO vehicular relay networks via cooperative jamming and signal superposition," *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 10732–10747, Dec. 2017.
- [19] J. Choi, "Minimum power multicast beamforming with superposition coding for multiresolution broadcast and application to NOMA systems," *IEEE Trans. Commun.*, vol. 63, no. 3, pp. 791–800, Mar. 2015.
- [20] D. Zhai, R. Zhang, L. Cai, B. Li, and Y. Jiang, "Energy-efficient user scheduling and power allocation for NOMA-based wireless networks with massive IoT devices," *IEEE Internet Things J.*, vol. 5, no. 3, pp. 1857–1868, Jun. 2018.
- [21] Y. Fu, Y. Chen, and C. W. Sung, "Distributed power control for the downlink of multi-cell noma systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 6207–6220, Sep. 2017.
- [22] X. Li, C. Li, and Y. Jin, "Dynamic resource allocation for transmit power minimization in OFDM-based NOMA systems," *IEEE Commun. Lett.*, vol. 20, no. 12, pp. 2558–2561, Dec. 2016.
- [23] F. Fang, H. Zhang, J. Cheng, and V. C. M. Leung, "Energy-efficient resource allocation for downlink non-orthogonal multiple access network," *IEEE Trans. Commun.*, vol. 64, no. 9, pp. 3722–3732, Sep. 2016.
- [24] J. Choi, "Joint rate and power allocation for NOMA with statistical CSI," *IEEE Trans. Commun.*, vol. 65, no. 10, pp. 4519–4528, Oct. 2017.
- [25] F. Fang, H. Zhang, J. Cheng, S. Roy, and V. C. M. Leung, "Joint user scheduling and power allocation optimization for energy efficient NOMA systems with imperfect CSI," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2874–2885, Dec. 2017.
- [26] Y. Wu and L. P. Qian, "Energy-efficient NOMA-enabled traffic offloading via dual-connectivity in small-cell networks," *IEEE Commun. Lett.*, vol. 21, no. 7, pp. 1605–1608, Jul. 2017.
- [27] B. Xu, Y. Chen, J. R. Carrion, and T. Zhang, "Resource allocation in energy-cooperation enabled two-tier NOMA HetNets toward green 5G," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2758–2770, Dec. 2017.
- [28] Y. Liu, X. Li, F. R. Yu, H. Ji, H. Zhang, and V. C. M. Leung, "Grouping and cooperating among access points in user-centric ultra-dense networks with non-orthogonal multiple access," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2295–2311, Oct. 2017.
- [29] A. E. Mostafa, Y. Zhou, and V. W. S. Wong, "Connectivity maximization for narrowband IoT systems with NOMA," in *Proc. IEEE ICC'17*, Paris, France, May 2017, pp. 1–6.
- [30] D. Zhai and R. Zhang, "Joint admission control and resource allocation for multi-carrier uplink NOMA networks," *IEEE Wireless Commun. Lett.*, vol. 7, no. 4, pp. 922–925, Dec. 2018.
- [31] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6010–6023, Aug. 2016.
- [32] S. Shi, L. Yang, and H. Zhu, "Outage balancing in downlink nonorthogonal multiple access with statistical channel state information," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4718–4731, Jul. 2016.
- [33] P. Xu, Y. Yuan, Z. Ding, X. Dai, and R. Schober, "On the outage performance of non-orthogonal multiple access with 1-bit feedback," *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 6716–6730, Oct. 2016.
- [34] 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.
- [35] P. Xu and K. Cumanan, "Optimal power allocation scheme for non-orthogonal multiple access with α -fairness," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2357–2369, Oct. 2017.
- [36] D. Wang, P. Ren, and J. Cheng, "Cooperative secure communication in two-hop buffer-aided networks," *IEEE Trans. Commun.*, vol. 66, no. 3, pp. 972–985, Mar. 2018.
- [37] Z. Ding, D. W. K. Ng, R. Schober, and H. V. Poor, "Delay minimization for NOMA-MEC offloading," *IEEE Signal Process. Lett.*, vol. 25, no. 12, pp. 1875–1879, Dec. 2018.
- [38] Y. Endo, Y. Kishiyama, and K. Higuchi, "Uplink non-orthogonal access with MMSE-SIC in the presence of inter-cell interference," in *Proc. IEEE ISWCS'2012*, Paris, France, Aug. 2012, pp. 261–265.
- [39] J. Bondy and U. Murty, *Graph Theory*. Berlin, Germany: Springer-Verlag, 2008.
- [40] M. R. Garey and D. S. Johnson, *Computer and Intractability: A Guide to the Theory of NP-Completeness*. San Francisco, CA, USA: Freeman, 1979.
- [41] R. D. Yates, "A framework for uplink power cellular radio systems," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 7, pp. 1341–1347, Sep. 1995.



Daosen Zhai (S'16–M'18) received the B.E. degree in telecommunication engineering from Shandong University, Weihai, China, in 2012, and the Ph.D. degree in communication and information systems from Xidian University, Xi'an, China, in 2017. He is currently an Assistant Professor with the School of Electronics and Information, Northwestern Polytechnical University, Xian, China. His research interests focus on radio resource management in NOMA networks, energy harvesting networks, and 5G wireless networks.



Ruonan Zhang (S'09–M'10) received the B.S. and M.Sc. degrees in electrical and electronics engineering from Xi'an Jiaotong University, Xi'an, China, in 2000 and 2003, respectively, and the Ph.D. degree in electrical and electronics engineering from the University of Victoria, Victoria, BC, Canada, in 2010.

He was an IC Architecture Engineer with Motorola Inc. and Freescale Semiconductor Inc., Tianjin, China, from 2003 to 2006. Since 2010, he has been with the Department of Communication Engineering, Northwestern Polytechnical University, Xi'an,

China, where he is currently a Professor. His current research interests include wireless channel measurement and modeling, architecture and protocol design of wireless networks, and satellite communications.

Dr. Zhang was a recipient of the New Century Excellent Talent Grant from the Ministry of Education of China and the Best Paper Award of IEEE NaNA 2016. He has served as a Local Arrangement Co-Chair for the IEEE/CIC International Conference on Communications in China in 2013 and as an Associate Editor for the *Journal of Communications and Networks*.



Lin Cai (S'00–M'06–SM'10) received the M.A.Sc. and Ph.D. degrees (awarded outstanding achievement in graduate studies) in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 2002 and 2005, respectively. Since 2005, she has been with the Department of Electrical and Computer Engineering, University of Victoria, and she is currently a Professor. Her research interests span several areas in communications and networking, with a focus on network protocol and architecture design supporting emerging multimedia

traffic and Internet of Things.

Prof. Cai was the recipient of the NSERC Discovery Accelerator Supplement Grants in 2010 and 2015, respectively, and the recipient of the Best Paper Awards of IEEE ICC 2008 and IEEE WCNC 2011. She has founded and chaired IEEE Victoria Section Vehicular Technology and Communications Joint Societies Chapter. She has been elected to serve the IEEE Vehicular Technology Society Board of Governors, 2019–2021. She has served as an Area Editor for the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, a Member of the Steering Committee of the IEEE TRANSACTIONS ON BIG DATA (TBD) and the IEEE TRANSACTIONS ON CLOUD COMPUTING (TCC), an Associate Editor for the IEEE INTERNET OF THINGS JOURNAL, the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, the *EURASIP Journal on Wireless Communications and Networking*, the *International Journal of Sensor Networks*, and the *Journal of Communications and Networks* (JCN), and as the Distinguished Lecturer of the IEEE VTS Society. She has served as a TPC symposium Co-Chair for IEEE Globecom'10 and Globecom'13. She is a registered Professional Engineer of British Columbia, Canada.



F. Richard Yu (S'00–M'04–SM'08–F'18) received the Ph.D. degree in electrical engineering from the University of British Columbia (UBC) in 2003. From 2002 to 2006, he was with Ericsson (in Lund, Sweden) and a start-up in California, USA. He joined Carleton University in 2007, where he is currently a Professor. His research interests include wireless cyber-physical systems, connected/autonomous vehicles, security, distributed ledger technology, and deep learning.

He serves on the editorial boards of several journals, including Co-Editor-in-Chief for *Ad Hoc & Sensor Wireless Networks*, Lead Series Editor for *IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY*, *IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING*, and *IEEE COMMUNICATIONS SURVEYS & TUTORIALS*. He has served as the Technical Program Committee (TPC) Co-Chair of numerous conferences. Dr. Yu is a registered Professional Engineer in the province of Ontario, Canada, a Fellow of the Institution of Engineering and Technology (IET), and a Fellow of the IEEE. He is a Distinguished Lecturer, the Vice President (Membership), and an elected member of the Board of Governors (BoG) of the IEEE Vehicular Technology Society. He received the IEEE Outstanding Service Award in 2016, IEEE Outstanding Leadership Award in 2013, Carleton Research Achievement Award in 2012, the Ontario Early Researcher Award (formerly Premier's Research Excellence Award) in 2011, the Excellent Contribution Award at IEEE/IFIP TrustCom 2010, the Leadership Opportunity Fund Award from Canada Foundation of Innovation in 2009 and the Best Paper Awards at IEEE ICNC 2018, VTC 2017 Spring, ICC 2014, Globecom 2012, IEEE/IFIP TrustCom 2009 and Int'l Conference on Networking 2005.