Distributed User Clustering and Resource Allocation for Imperfect NOMA in Heterogeneous Networks

Distributed User Clustering and Resource Allocation for Imperfect NOMA in Heterogeneous Networks

Abdulkadir Celik, *Member, IEEE,* Ming-Cheng Tsai, *Student Member, IEEE,*
Redha M. Radaydeh, *Senior Member, IEEE,* Fawaz S. Al-Qahtani, *Member, IEEE,*
and Mohamed-Slim Alouini, *Fellow, IEEE*

***Abstract*—In this paper, we propose a distributed cluster formation (CF) and resource allocation (RA) framework for non-ideal non-orthogonal multiple access (NOMA) schemes in heterogeneous networks. The imperfection of the underlying NOMA scheme is due to the receiver sensitivity and interference residue from non-ideal successive interference cancellation (SIC), which is generally characterized by a fractional error factor (FEF). Our analytical findings first show that several factors have a significant impact on the achievable NOMA gain. Then, we investigate fundamental limits on NOMA cluster size as a function of FEF levels, cluster bandwidth, and quality of service (QoS) demands of user equipments (UEs). Thereafter, a clustering algorithm is developed by taking feasible cluster size and channel gain disparity of UEs into account. Finally, we develop a distributed α-fair RA framework where α governs the trade-off between maximum throughput and proportional fairness objectives. Based on the derived closed-form optimal power levels, the proposed distributed solution iteratively updates bandwidths, clusters, and UEs' transmission powers. Numerical results demonstrate that proposed solutions deliver a higher spectral and energy efficiency than traditionally adopted basic NOMA cluster size of two. We also show that an imperfect NOMA cannot always provide better performance than orthogonal multiple access under certain conditions. Finally, our numerical investigations reveal that NOMA gain is maximized under downlink/uplink decoupled (DUDe) UE association.**

***Index Terms*—Downlink uplink decoupling, alpha fairness, proportional fairness, imperfect SIC, residual interference.**

I. Introduction

EVER increasing number of communications devices with the ambitious quality of service (QoS) demands puts forward challenging goals for fifth-generation (5G) networks such as massive connectivity, enhanced mobile broadband, ultra-reliability, low-latency, etc. To fulfill such demands, ultra-dense heterogeneous networks (HetNets) have already been considered as a promising solution since densification of the network has the ability to boost the network coverage and capacity while reducing the operational and capital expenditures of mobile operators [1]. However, traditional orthogonal multiple access (OMA) schemes employed by today's HetNets dedicate radio resources to a certain user either in time, frequency, or code domains, which is not adequately spectral efficient for the expected massive number of users. Having its root in multi-user detection theory, non-orthogonal multiple access (NOMA) can momentarily serve multiple users on the same radio resource by multiplexing them either in power or code domain [2]. As a result, it has recently gained attention with its ability to serve toward higher spectral efficiency and massive connectivity goals of the next generation networks. In particular, power domain NOMA ensures a certain reception power for each user such that some users operate in low power levels in order to cancel dominant interference using successive interference cancellation (SIC) while some others transmit at high power levels at the expense of limited interference cancellation (IC) opportunity. Even though such a strategy paves the way for a notion of fairness embedded in NOMA, the impacts of fair bandwidth scheduling on the network performance is still an interesting phenomenon to be investigated.

In order to extract the desired signal, the SIC receiver first decodes the strongest interference, then re-generates the transmitted signal, and finally subtracts it from the received composite signal, which is repeated for succeeding interference components. However, a real-life NOMA system is required to account for the following challenges of a practical SIC receiver: First, a more complicated power control policy is necessary since decoder needs to observe a minimum SINR at each cancellation stage, which is mainly characterized by the receiver sensitivity [3]. Thus, the optimal power control strategy must comply with the resulting disparity of the received power levels, which is also referred to as *SIC constraints*. Second, system performance can substantially be deteriorated due to the amplitude and phase estimation errors which determine the residual interference after SIC and is often quantified by a *fractional error factor* (FEF) [4]. Therefore, it is necessary to develop an optimal power and bandwidth allocation scheme which accounts for these practical challenges.

Furthermore, cluster formation (CF) is of the utmost importance to maximize the gain achieved by a NOMA scheme. Ongoing research efforts typically consider a perfect NOMA scheme for basic cluster of size two, which directly reduces the clustering to a pairing problem. However, clustering more user equipments (UEs) to share the same bandwidth provides an improved spectral efficiency at the cost of SIC delay which linearly increases with the cluster size. Therefore, CF problem

A. Celik, Ming-Cheng Tsai, and M-S. Alouini are with Computer, Electrical, and Mathematical Sciences and Engineering Division at King Abdullah University of Science and Technology (KAUST), Thuwal, KSA. Corresponding author: Abdulkadir Celik (abdulkadir.celik@kaust.edu.sa).

R. M. Radaydeh is with the Electrical Engineering Program, Department of Engineering and Technology, Texas A& M University-Commerce, , Commerce, TX 75428 USA.

F. S. Al-Qahtani is with Research, Development, Innovation (RDI) Division of Qatar Foundation, Doha, Qatar.

This work was supported by NPRP from the Qatar National Research Fund under grant no. 8-1545-2-657. A part of this paper has been presented at IEEE GLOBECOM 2018, Singapore [1].

involves two main tasks: 1) Determining the optimal cluster sizes and 2) Grouping UEs to maximize the overall network performance. Accordingly, this paper addresses a distributed framework for cluster formation (CF) and α-fair resource allocation for UL-HetNets under imperfect NOMA scheme.

A. Related Works

Related works on NOMA can be exemplified as follows: The impact of UE grouping is investigated in [5] for a two-UE DL-NOMA system with fixed and cognitive radio inspired power allocation schemes. The work in [6] proposed three different sub-optimal approaches for max-min fair UE clustering problem. Authors of [7] iteratively built clusters where each iteration jointly optimizes beam-forming and power allocation for given clusters. User pairing for UL-NOMA is investigated in [8] which divides the set of UEs into disjunct pairs and assigns the available resources to these pairs by considering some predefined power allocation schemes. In [9], authors study the optimal user pairing for the NOMA system in the sense of maximizing the total sum rate. In [10], authors study the problem of resource optimization, mode selection, and power allocation subject to queue stability constraints under the assumption of in-band full duplex base stations (BSs). Beam-forming and power allocation of a multiple-input-multiple-output (MIMO) NOMA system are investigated in [11] where two-UE clusters are formed from high and low channel gain UEs with the consideration of channel gain correlations. The work in [12] proposed a UL power allocation scheme by first grouping users into a single cluster and then optimizing the power allocation.

In [13], Ding et. al. proposed a cluster beamforming strategy to jointly optimize beamforming vectors and power allocation coefficients for MIMO-NOMA clustering with the purpose of energy efficiency. The sum rate maximization problem of mm-wave-NOMA systems is investigated in [14] where authors also develop a K-means-based machine learning algorithm for user clustering. In [15], a suboptimal quality-balanced clustering approach is proposed to optimize the total sum rate in a system. The impacts of channel state information (CSI) imperfections on the NOMA performance are investigated in [16]–[18]: Energy efficient resource scheduling is addressed in [16] where authors also account for imperfect CSI for a generic cluster size. In [17], power-efficient resource allocation is studied for multicarrier NOMA systems. Accounting for the imperfection of CSI at transmitter side, a solution is proposed to jointly design the power-rate allocation, user scheduling, and SIC decoding policy for minimizing the total transmit power. An interesting problem is investigated in [18] for NOMA systems vulnerable to jamming attacks. Authors proposed a reinforcement learning-based power control scheme without being aware of the jamming and CSI. Except [11]–[16], proposed methods in these works mostly focus on the basic form of a NOMA (i.e., pairing only two UEs where power allocation is analytically more tractable) by ignoring the benefits of incorporating more UEs.

Considering the massive connectivity goal of the future networks, it is important to investigate NOMA schemes that allow larger cluster sizes for the sake of spectral efficiency and increased connectivity. Since it is possible to employ sophisticated SIC receivers at BSs with high computational power, possible IC delay of larger cluster sizes can be mitigated in order to enhance UL-NOMA performance.

In [19], multi-cell uplink NOMA systems are analyzed using the theory of Poisson cluster process. The impact of channel gain disparity on DL-NOMA is investigated in [5] for a two-user system with fixed and cognitive radio inspired power allocation. A near optimal solution was proposed by combining Lagrangian duality and dynamic programming for joint power and channel allocation in [20]. In [21], the authors derived closed-form expressions for the outage probability of two-user UL-NOMA assuming fixed powers of different users. A simple power and rate allocation scheme for UL-NOMA is developed for a multicarrier system in [22] where a practical modulation and coding scheme is employed at each UE. In [23], a distributed UL-NOMA scheme is proposed for cloud radio access networks, which can offer substantial improvement over benchmark schemes. In these works, authors mostly consider a basic NOMA cluster of size two except [12] where authors develop a general DL and UL power control framework for a generic cluster size.

In [24], a dynamic power allocation scheme is proposed for both DL and UL NOMA scenarios with two users with various QoS requirements. User clustering and power-bandwidth allocation of HetNets is studied in [1], [25] where clusters are formed according to different objectives such as maximum sum-rate, max-min fairness, and energy-spectrum cost minimization. The work in [25] is further extended for for DL-HetNets in [26] where a user clustering and power-bandwidth allocation is proposed. The main limitation of this work is treating the cluster size as a given design parameter. However, it is necessary to analyze the maximum permissible cluster size for UL-HetNets since the next-generation networks are expected to accommodate the massive connectivity. Therefore, allowing more low-power users on the same resource block is desirable for serving a large number of users and increasing the spectral efficiency of the NOMA scheme. In this regard, the proposed user clustering in this paper is apart from that of [26] such that we analytically derive the maximum cluster size as a function of QoS demands, channel quality, cluster bandwidth, and SIC efficiency. Accordingly, the proposed clustering algorithm forms the clusters jointly with bandwidth allocation and ensures the QoS demand of each cluster is satisfied. Noting that [26] is not involved with resource allocation fairness, our closed-form power allocations are also different since UL-UEs do not compete for a common power source. Excluding [1], [25], [26], these works also do not consider the residual interference caused by the error propagation during the IC process.

B. Main Contributions

Our main contributions can be summarized as follows:

- An imperfect NOMA scheme is investigated in order to account for practical SIC constraints due to the receiver sensitivity and residual interference. Our analytical find-

ings show that decoding order, SIC constraints, residual interference, and channel gain disparity of cluster members have a significant impact on the achievable NOMA gain. These findings are then supported with numerical results which clearly demonstrate that NOMA cannot always provide a better performance than OMA depending upon the SINR requirements and FEF levels (i.e, residual interference).

- Existing works on NOMA typically assume a basic cluster size of two and simply pair UEs to form clusters. However, cluster sum-rates and spectral efficiency increase with the cluster size as more UEs share the same bandwidth. Hence, the largest feasible cluster size is first analytically obtained as a function of FEF levels, cluster bandwidth, and QoS requirements of cluster members. Then, we show that a given bandwidth can accommodate more UEs as the SINR requirements and FEF levels decrease, which is especially beneficial to outline capabilities of NOMA to serve for the massive connectivity. Thereafter, we propose a cluster formation method where each BS first determines the largest feasible cluster sizes and then iteratively matches its UEs with clusters to maximize the channel gain disparity for an improved NOMA gain.
- Based on the developed cluster formation method, we develop a distributed α-fair resource allocation methodology where $\alpha \in [0,1]$ manages the balance between maximum throughput and proportional fairness. Resource allocation is decomposed into power control and bandwidth allocation problems. Based on the derived closed-form power control expression of the considered imperfect NOMA scheme, the proposed algorithm iteratively updates bandwidth allocations, cluster sizes, and transmission powers to maximize the α-fair network objective. Finally, the performance gain of the developed algorithm is compared with OMA and basic NOMA schemes under different system network parameters such as BS/UE density, traffic offloading factor, FEF, and QoS requirements.

C. Notations and Paper Organization

Throughout the paper, sets and their cardinality are denoted with calligraphic and regular uppercase letters (e.g., $|\mathcal{A}| = A$), respectively. Vectors and matrices are represented in lowercase and uppercase boldfaces (e.g., $\boldsymbol{a}$ and $\boldsymbol{A}$), respectively. Superscripts b, c, and i are used for indexing BSs/cells, clusters, and UEs, respectively. The remainder of the paper is organized as follows: Section II introduces the network model and UE association schemes. Section III addresses constraints, decoding order, and imperfections of practical SIC receivers and their impacts on achievable NOMA gain. Section IV first provides the problem statement and then gives an overview of proposed solution methodology. Section V analyses the feasible cluster size and develops a clustering algorithm. Section VI addresses proposed α-fair distributed power and bandwidth allocation along with the algorithm of overall solution. Numerical results are presented in Section VII and Section VIII concludes the paper with a few remarks.

II. Network Model

We consider UL transmission of a 2-tiered HetNet that operates on a single-input and single-output NOMA scheme. Each tier represents a particular cell class, i.e., tier-1 consists of a single macrocell and tier-2 comprises of smallcells. The index set of all BSs is denoted by $\mathcal{B} = \{b \,|\, 0 \leqslant b \leqslant B\}$ where B denotes the number of small base staions (SBSs), $b = 0$ is for the macro base station (MBS) index, and $1 \leqslant b \leqslant B$ are indices for SBSs, respectively[1]. Maximum transmission powers of UEs and BSs are denoted as $\bar{P}_u$ and $\bar{P}_c$, respectively, where $\bar{P}_c$ equals to $\bar{P}_m$ and $\bar{P}_s$ for the MBS and SBSs, respectively.

Association of UEs with the BS can be done either in a DL/UL coupled (DUCo) or decoupled (DUDe) fashion. Conventional DUCo scheme associates UEs with the same BS for both DL and UL transmission based on received signal strength information (RSSI), which yields a significant traffic load on macrocells due to MBS's high transmission power. Therefore, UE association is typically done by introducing a bias factor, $0 \leqslant \flat \leqslant 1$, in order to offload DL traffic from MBSs to SBSs. Nonetheless, requiring UEs to follow the same association in both UL and DL may always not yield a desirable performance. While eeping DL association method the same as in DUCo, DUDe scheme alternatively determines the UL association based on channel gain such that a UE can be associated with a nearby SBS in the UL even if it is associated with the MBS in the DL [27], [28].

Contingent upon the user associations, index set of all $U \triangleq \sum_b U_b$ UEs is given as $\mathcal{U} \triangleq \bigcup_b \mathcal{U}_b$ where $\mathcal{U}_b$ is the set of U_b UEs associated with BS$_b$. Each BS partitions $\mathcal{U}_b$ into disjoint C_b clusters such that $\mathcal{K}_b^c$ symbolizes the set of K_b^c UEs within cluster c, i.e., $U_b = \sum_{c \in \mathcal{C}_b} K_b^c$. Similarly, the set of all C clusters are denoted as $\mathcal{C} \triangleq \bigcup_b \mathcal{C}_b$ where $\mathcal{C}_b$ is the set of C_b clusters of BS$_b$. Entire UL bandwidth is divided into Θ resource blocks (RBs) each of which has a bandwidth of W Hz. The available set of RBs can be exploited by C clusters based on an α-fair resource allocation policy. The number of RBs allocated to $\mathcal{K}_b^c$ is denoted as $\theta_b^c \in [0, \Theta]$, $\sum_{b,c} \theta_b^c \leqslant \Theta$. For the remainder of the paper, we assume that a UE can be associated with exactly one cluster at a time and allocated RBs are dedicated to the corresponding clusters.

III. Impacts of Constraints on NOMA Gain

In this section, we first introduce the constraints and imperfections of a practical SIC receiver, then analyze the impacts of decoding order and receiver sensitivity on the achievable NOMA gain.

A. Constraints and Imperfections of SIC Receivers

Let us now focus on a generic cluster of BS$_b$ $\mathcal{K}_b^c = \{i |\, i \in \mathcal{U}_b, h_{i-1}^b \geqslant h_i^b \geqslant h_{i+1}^b, \delta_{b,c}^i = 1\}$ where $\delta_{b,c}^i \in \{0,1\}$ is a

[1]The terms BS, cell, and their indices are used interchangeably throughout the paper.

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/TCOMM.2019.2927561, IEEE Transactions on Communications

binary indicator for cluster membership. For the UL-NOMA transmission, we consider the following decoding order

$$\underbrace{\omega_{b,c}^{K_b^c} h_{K_b^c}^b < ... <}_{\substack{\text{Lower Rank Decoding Order} \\ \mathcal{O}_i^\ell \text{ that can not be cancelled}}} \quad \omega_{b,c}^i h_i^b \quad \underbrace{< ... < \omega_{b,c}^1 h_1^b}_{\substack{\text{Higher Rank Decoding Order} \\ \mathcal{O}_i^h \text{ that can be cancelled}}} \quad , \tag{1}$$

where h_i^b is the composite channel gain from UE_i to BS_b, $\omega_{b,c}^i h_i^b$ is the power received from UE_i which is normalized by the maximum transmission power $\bar{P}_u$, $\omega_{b,c}^i$ is the power allocation weight, $\mathcal{O}_i^\ell = \{i+1, \ldots, K_b^c\}$ is the lower rank decoding order set, and $\mathcal{O}_i^h = \{1, \ldots, i-1\}$ is the higher rank decoding order set for UE_i. Notice that the UL decoding order is the reverse of DL order considered in the literature, which is addressed in the next section. Accordingly, a generic SINR representation of the imperfect SIC receiver can be given by

$$\Gamma_{b,c}^i = \frac{\delta_{b,c}^i \omega_{b,c}^i h_i^b}{\epsilon_b \sum_{\substack{j=1 \\ j \in \mathcal{K}_b^c}}^{i-1} \delta_{b,c}^j \omega_{b,c}^j h_j^b + \sum_{\substack{k=i+1 \\ j \in \mathcal{K}_b^c}}^{K_b^c} \delta_{b,c}^k \omega_{b,c}^k h_k^b + \varrho_b^c}, \tag{2}$$

where $0 \leqslant \epsilon_b \leqslant 1$ is the FEF of BS_b which characterizes the residual interference, $\varrho_b^c \triangleq \sigma \theta_b^c / \bar{P}_u$, $\sigma = N_0 \theta_b^c W$ is the thermal receiver noise power, and N_0 is the noise power spectral density. The first term of the denominator represents the residual interference after cancellation which can indeed be linked to the SIC efficiency, i.e., $(1-\epsilon_b)$. On the other hand, the second term of denominator represents the uncancelled lower rank interference.

The first term of the denominator represent the residual interference after cancellation which can indeed be linked to the SIC efficiency, i.e., $(1-\epsilon_b)$. On the other hand, the second term of denominator represents the uncancelled lower rank interference. The residual interference is primarily caused by amplitude, phase, and channel estimation errors, which lead to imperfect regeneration of the received signals. Another source of SIC imperfections is erroneous bit decisions in the previously decoded users. Under low bit-error rate requirements ($< 10^{-5}$), error propagation of bit decisions is also a result of the imperfect estimations [29]. Multistage detection, error correction coding, iterative detection, enhanced channel estimation are among the key techniques to ameliorate FEF levels of SIC receivers [30]. That is, the FEF is an important hardware parameter to be taken into account in power allocation strategy because being FEF agnostic can substantially deteriorate the NOMA performance as $\epsilon_b \to 1$.

For a desirable performance, a cluster member should be able to cancel the dominant interference while tolerating the SIC imperfection and interference induced from lower rank UEs. Following from (2), the achievable data rate of UE_i is given by

$$\mathrm{R}_i = W \theta_b^c \log_2(1 + \Gamma_{b,c}^i), \forall i \in \mathcal{K}_b^c. \tag{3}$$

R_i is generally required to be higher than a certain service rate agreement, $\mathrm{R}_i \geqslant \overline{\mathrm{R}}_i, \forall i$, which is referred to as *QoS constraint*[2] and given by

$$\Gamma_{b,c}^i \geqslant 2^{\frac{\overline{\mathrm{R}}_i}{\theta_b^c W}} - 1, \forall i \in \mathcal{K}_b^c, \forall b, \forall c. \tag{4}$$

On the other hand, SIC constraints are given by

$$\Gamma_{b,c}^i \geqslant 10^{\frac{\S_b}{10}}, \forall i \in \mathcal{K}_b^c, \forall b, \forall c, \tag{5}$$

where $\S_b$ is the receiver sensitivity of BS_b which is often given in units of dB. These two constraints can be combined and projected onto SINRs as a unified constraint as follows

$$\Gamma_{b,c}^i \geqslant \bar{\Gamma}_{b,c}^i(\theta_b^c) = \max\left(10^{\frac{\S_b}{10}}, 2^{\frac{\overline{\mathrm{R}}_i}{\theta_b^c W}} - 1\right), \forall i \in \mathcal{K}_b^c, \tag{6}$$

which is referred to as *composite SINR constraints* (CSCs) in the remainder of paper.

B. Impacts of SIC Constraints and Imperfections

Even though DL-NOMA decodes UE signals in descending order of their channel gains, employing the same order in UL-NOMA may not give the desired performance under CSCs. To be more specific, let us consider a basic NOMA cluster of UE_k and UE_ℓ with channel gains h_k and h_ℓ, $h_k \geqslant h_\ell$, and composite SINR demands $\bar{\Gamma}_k$ and $\bar{\Gamma}_\ell$, respectively.

1) Descending Order: Employing the descending decoding order as in the DL case (i.e., UE_k cancel the interference of UE_ℓ), OMA and NOMA sum-rates can be respectively given as

$$\mathrm{R}_\downarrow^O = {}^1\!/_2 \left\{\log_2(1+\rho h_k) + \log_2(1+\rho h_\ell)\right\}, \tag{7}$$

$$\mathrm{R}_\downarrow^N = \log_2\left(1 + \frac{\omega_k h_k}{\epsilon \omega_\ell h_\ell + {}^1\!/_\rho}\right) + \log_2\left(1 + \frac{\omega_\ell h_\ell}{\omega_k h_k + {}^1\!/_\rho}\right), \tag{8}$$

where $0 \leqslant \omega_k, \omega_\ell \leqslant 1$ are power weights and $\rho = {}^{\bar{P}_u}\!/_\sigma$. As $\rho \to \infty$ and $\epsilon \to 0$, asymptotic capacity of OMA and perfect NOMA can be respectively expressed as

$$\lim_{\substack{\rho \to \infty \\ \epsilon \to 0}} \mathrm{R}_\downarrow^O \simeq {}^1\!/_2 \{\log_2(\rho h_k) + \log_2(\rho h_\ell)\} = {}^1\!/_2 \log_2\left(\rho^2 h_k h_\ell\right), \tag{9}$$

$$\lim_{\substack{\rho \to \infty \\ \epsilon \to 0}} \mathrm{R}_\downarrow^N \simeq \log_2(\rho \omega_k h_k) + \log_2\left(1 + \frac{\omega_\ell h_\ell}{\omega_k h_k}\right) \simeq \log_2(\rho \omega_k h_k), \tag{10}$$

where (9) and (10) follow from the facts that $(1+\rho h_k) \simeq \rho h_k$ as $\rho \to \infty$ and the second term of (8) becomes negligible as $\rho \to \infty$, respectively. Accordingly, asymptotic gain of NOMA scheme can be given by

$$\begin{aligned} \Delta_\downarrow &\triangleq \lim_{\substack{\rho \to \infty \\ \epsilon \to 0}} \left(\mathrm{R}_\downarrow^N - \mathrm{R}_\downarrow^O\right) = \log_2(\rho \omega_k h_k) - {}^1\!/_2 \log_2\left(\rho^2 h_k h_\ell\right) \\ &= \log_2\left(\frac{\rho \omega_k h_k}{\rho \sqrt{h_k h_\ell}}\right) = \log_2\left(\omega_k \sqrt{\frac{h_k}{h_\ell}}\right) \end{aligned} \tag{11}$$

In the descending order, the SIC constraint requires $\lim_{\rho \to \infty} \frac{\omega_\ell h_\ell}{\omega_k h_k + {}^1\!/_\rho} \geqslant 10^{\frac{\S_b}{10}}$ that reduces to a power disparity

[2]Instead of the inelastic traffic conditions where users require a minimum instantaneous throughput requirements, we are interested in elastic users with average QoS demands over a long time period.

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/TCOMM.2019.2927561, IEEE Transactions on Communications

constraint, i.e., $\frac{\omega_\ell h_\ell}{\omega_k h_k} \geqslant 10^{\frac{\S_b}{10}}$. Even for a SIC receiver with perfect sensitivity, i.e., $\S_b \to 0$, power disparity constraint constitutes $\frac{\omega_\ell h_\ell}{h_k} \geqslant \omega_k$, thus the upper bound on $\Delta_\downarrow$ is given by

$$\Delta_\downarrow \leqslant \log_2\left(\omega_\ell \sqrt{\frac{h_\ell}{h_k}}\right) \leqslant {}^1\!/_2 \left\{\log_2(h_\ell) - \log_2(h_k)\right\}, \quad (12)$$

which is always non-positive due to ${}^{h_\ell}\!/_{h_k} \leqslant 1$. That is, sum-rate of UL-OMA and the descending ordered UL-NOMA perform the same for users with equal channel gains. For non-equal channel gain cases, UL-NOMA provides a worse performance which is deteriorated even further for imperfect NOMA case as $\epsilon \to 1$.

2) Ascending Order: Following the similar steps in (7)-(10), asymptotic NOMA gain for the ascending order case (i.e., UE_ℓ cancel the interference of UE_k) can be obtained as

$$\begin{aligned}\Delta_\uparrow &\triangleq \lim_{\substack{\rho\to\infty \\ \epsilon\to 0}} \left(\mathrm{R}_\uparrow^N - \mathrm{R}_\uparrow^O\right) = \log_2\left(\rho\omega_\ell h_\ell\right) - {}^1\!/_2 \log_2\left(\rho^2 h_k h_\ell\right) \\ &= \log_2\left(\frac{\rho\omega_\ell h_\ell}{\rho\sqrt{h_k h_\ell}}\right) = \log_2\left(\omega_\ell\sqrt{\frac{h_\ell}{h_k}}\right). \end{aligned} \quad (13)$$

In the ascending order, the SIC constraint requires $\lim_{\rho\to\infty} \frac{\omega_k h_k}{\omega_\ell h_\ell + {}^1\!/_\rho} \geqslant 10^{\frac{\S_b}{10}}$ that reduces to a power disparity constraint, i.e., $\frac{\omega_k h_k}{\omega_\ell h_\ell} \geqslant 10^{\frac{\S_b}{10}}$. For a SIC receiver with perfect sensitivity, i.e., $\S_b \to 0$, power disparity constraint constitutes $\frac{\omega_k h_k}{h_\ell} \geqslant \omega_\ell$, thus the upper bound on $\Delta_\downarrow$ is given by

$$\Delta_\uparrow \leqslant \log_2\left(\omega_k \sqrt{\frac{h_k}{h_\ell}}\right) \leqslant {}^1\!/_2 \left\{\log_2(h_k) - \log_2(h_\ell)\right\}, \quad (14)$$

which is always non-negative due to ${}^{h_k}\!/_{h_\ell} \geqslant 1$. That is, sum-rate of UL-OMA and the descending ordered UL-NOMA perform the same for users with equal channel gains whereas UL-NOMA provides a superior performance proportional to the channel gain disparity of users. Unfortunately, this desirable performance gain obtained by channel gain disparity of users naturally diminishes as ϵ increases and NOMA yields a worse performance than OMA after a certain point, which is investigated in the remainder of the paper.

IV. Cluster formation and Resource Allocation

Centralized CF and RA is a combinatorial problem whose solution requires impractical time complexity even for moderate size of HetNets. Since a fast yet high performance solution is of the essence to employ NOMA in large-scale HetNets, this section first makes a problem statement by formulating a centralized problem then outlines the proposed distributed solution methodology to mitigate the high communication and computational overhead of centralized solutions.

A. Centralized Problem Formulation

In order to investigate fair power and bandwidth allocation schemes, we adopt a generalized throughput formulation that has been proposed by the nominal work in [31] where the degree of fairness is adjusted by a single parameter $\alpha \in [0, 1]$. In other words, α manages the compromise between throughput maximization and fairness by means of the generalized α-fair function which can be expressed as

$$\pi_{b,c}^i = \begin{cases} \frac{1}{1-\alpha} R_i^{1-\alpha}\left(\delta_{b,c}^i, \theta_b^c, \omega_{b,c}^i\right), & \text{for } 0 \leqslant \alpha < 1 \\ \log\left[R_i\left(\delta_{b,c}^i, \theta_b^c, \omega_{b,c}^i\right)\right], & \text{for } \alpha = 1 \end{cases}, \quad (15)$$

which corresponds to the throughput maximization if $\alpha = 0$ and proportional fairness if $\alpha \to 1$. For the sake of a unified and continuous form of the fairness function, we exploit the following α-fair objective function [32]

$$\begin{aligned}\Pi\left(\boldsymbol{\delta}, \boldsymbol{\theta}, \boldsymbol{\omega}\right) &= \sum_{\forall(b,c,i)} \pi_{b,c}^i \left(\delta_{b,c}^i, \theta_b^c, \omega_{b,c}^i\right) \\ &= \sum_{\forall(b,c,i)} \frac{1}{1-\alpha}\left(R_i^{1-\alpha}\left(\delta_{b,c}^i, \theta_b^c, \omega_{b,c}^i\right) - 1\right). \end{aligned} \quad (16)$$

Accordingly, a centralized CF and RA problem can be formulated as in $\mathrm{P_o}$ where C_o^1 ensures that UEs are assigned to exactly one cluster and C_o^2 limits the number of UEs within a cluster by K_b^c. C_o^3 constraints the total number of RB allocation to available number of RBs, Θ. The power weight limitation on UE_i is introduced in C_o^4 where the power allocation for UE_i on cluster c is set to zero if $\text{UE}_i \notin \mathcal{K}_b^c$. CSCs are given by C_o^5 in order to account for QoS and SIC constraints. Finally, C_o^6 indicates the variable domains.

$$\begin{aligned} \mathrm{P_o}: \max_{\boldsymbol{\delta},\boldsymbol{\theta},\boldsymbol{\omega}} \quad & \Pi(\boldsymbol{\delta}, \boldsymbol{\theta}, \boldsymbol{\omega}) \\ \mathrm{C}_o^1: \quad \text{s.t.} \quad & \sum_c \delta_{b,c}^i = 1, \qquad \forall b, i \\ \mathrm{C}_o^2: \quad & \sum_i \delta_{b,c}^i \leqslant K_b^c, \qquad \forall b, c \\ \mathrm{C}_o^3: \quad & \sum_{b,c} \theta_b^c \leqslant \Theta, \\ \mathrm{C}_o^4: \quad & \omega_{b,c}^i \leqslant \delta_{b,c}^i, \qquad \forall b, c, i \\ \mathrm{C}_o^5: \quad & \bar{\Gamma}_{b,c}^i(\theta_b^c) \leqslant \Gamma_{b,c}^i, \qquad \forall b, c, i \\ \mathrm{C}_o^6: \quad & \delta_{b,c}^i \in \{0,1\}, K_b^c \in [0, {}^U\!/_2], \theta_b^c \in [0,1] \end{aligned} \quad (17)$$

B. Hierarchically Distributed Solution

In $\mathrm{P_o}$, obtaining optimal integer valued cluster sizes and binary valued UE-cluster associations yields an NP-Hard mixed integer non-linear programming (MINLP) problem whose time complexity exponentially increases with the number of network entities. Moreover, highly non-convex nature of resource allocation problem puts an additional degree of complexity. Also noting the undesirable communication overhead of centralized solutions, developing fast yet near optimal distributed solutions is of interest to be employed in practice.

As shown in Fig. 1, we develop a distributed solution methodology where we first decouple the CF and power allocation problems by considering the channel gain disparity of cluster members as the main credential of cluster formation policy. This is primarily motivated by the analytical findings of Section III which shows that NOMA gain is determined by the channel gain disparity of the cluster members. In this way, each BS can independently form its own clusters since they

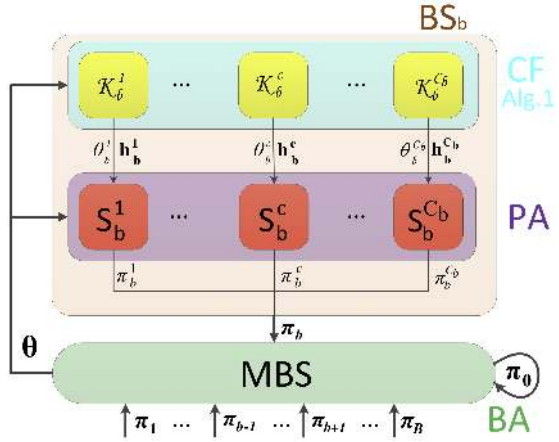


Fig. 1: Illustration of the proposed distributed clustering and resource allocation scheme [c.f. Algorithm 2]

are generally aware of the channel states of associated UEs. Notice that the CF problem is still coupled by the bandwidth allocations since the maximum permissible cluster size is a function of the cluster bandwidth as explained in the next section.

On the other hand, resource allocation problem is further decomposed into slave and master problems which are responsible for power and bandwidth allocation, respectively. Given cluster members and bandwidths, each slave problem is accountable for obtaining an optimal power control policy for imperfect NOMA scheme subject to CSCs. Thereafter, achieved cluster utilities are shared by a central unit (preferably the MBS) that accordingly updates the cluster bandwidths for the next iteration, which is followed by another round of cluster formation and power allocation, so on and so forth. The details of the proposed distributed solution methodology are addressed in the following sections.

V. Design and Analysis of NOMA Clusters

NOMA clustering involves two main design tasks; 1) determining the number of clusters and their size and 2) assigning UEs to the clusters. Accordingly, this section first analyzes the cluster size based on random matrix theory and derives the largest feasible cluster size as a closed-form function of the FEF levels, CSCs, and cluster bandwidth. Then, we propose a weighted maximum matching based CF method by weighting edges with channel gain disparity of UEs.

A. Fundamental Limits of NOMA Clusters

Without loss of generality, let us consider a cluster of size K and bandwidth allocation θ, whose CSCs can be written in the matrix form as

$$(\mathbf{I}-\boldsymbol{\Gamma}(\theta)\mathbf{H})\,\mathbf{p} \geqslant \bar{\boldsymbol{\Gamma}}(\theta)\boldsymbol{\sigma},\ \text{s.t. } \mathbf{p}>\mathbf{0}, \tag{18}$$

where vectors are of size $1\times K$, matrices are of size $K\times K$, $\mathbf{I}$ is the identity matrix, $\bar{\boldsymbol{\Gamma}}(\theta) = \mathrm{diag}(\bar{\Gamma}_1(\theta),\ldots,\bar{\Gamma}_k(\theta),\ldots,\bar{\Gamma}_K(\theta))$ is the diagonal matrix of the composite SINR demands, $\mathbf{p}$ is the column vector of the received powers, $\boldsymbol{\sigma}$ is the column vector of the receiver noise, and $\mathbf{H}$ is the interference channel gain matrix with entries

$$H_i^j = \begin{cases} 1, & i<j \\ 0, & i=j \\ \epsilon, & i>j \end{cases}, \tag{19}$$

where cases correspond to uncancelled interference, self interference, and residual interference, respectively. Notice that $\mathbf{H}$ has non-negative elements and is generally considered to be irreducible [33][3]. For a non-negative irreducible matrix, *Perron-Frobenius theorem* teaches us that the maximum eigenvalue of $\mathbf{H}$ is real-positive and eigenvector corresponding to the maximum eigenvalue is non-negative [35]. Following from the facts known from the standard matrix theory, a necessary and sufficient condition for the existence of a feasible solution to (18) requires the magnitude of the maximum eigenvalue of $\mathbf{F}\triangleq\bar{\boldsymbol{\Gamma}}(\theta)\mathbf{H}$ to be less than unity, i.e., $\lambda_F<1$ [33]. Assuming the existence of a feasible solution, a Pareto-optimal solution to (18) is then given by $\mathbf{p}^* = (\mathbf{I}-\boldsymbol{\Gamma}(\theta)\mathbf{H})^{-1}\bar{\boldsymbol{\Gamma}}(\theta)\boldsymbol{\sigma}$ where any other feasible $\mathbf{p}$ satisfying (18) would require more power than $\mathbf{p}^*$, i.e., $\mathbf{p}\geqslant\mathbf{p}^*$. From energy efficiency point of view, we stick with the minimum power consuming solution $\mathbf{p}^*$. Based on these discussion, we introduce following lemmas for the largest feasible cluster size as a function of the FEF and CSCs.

Lemma 1 (Energy unconstrained cluster size)**.** *For a cluster of energy unconstrained UEs, the largest feasible cluster size falls within the range of* $K_{\min}(\epsilon,\theta)\leqslant K(\epsilon,\theta)\leqslant K_{\max}(\epsilon,\theta)$*, i.e.,*

$$\left\lceil \frac{\ln(\epsilon)}{\ln\left(\frac{1+\epsilon\max_i(\bar{\Gamma}_i(\theta))}{1+\max_i(\bar{\Gamma}_i(\theta))}\right)}\right\rceil \leqslant K(\epsilon,\theta)\leqslant \left\lfloor \frac{\ln(\epsilon)}{\ln\left(\frac{1+\epsilon\min_i(\bar{\Gamma}_i(\theta))}{1+\min_i(\bar{\Gamma}_i(\theta))}\right)}\right\rfloor. \tag{20}$$

Accordingly, $K^\star(\epsilon,\theta)=K_{\min}(\epsilon,\theta)$ *is the largest feasible cluster size which is mainly determined by the user with the highest composite SINR demand.*

Proof. Please see Appendix A. □

Lemma 2 (Cluster size for identical CSCs)**.** *As a special case,* $\bar{\Gamma}_i(\theta)=\bar{\Gamma}(\theta),\forall i$*, the range in* (20) *tightens to an exact size of* $K^*(\epsilon,\theta)=\left\lfloor\frac{\ln(\epsilon)}{\ln\left(\frac{1+\epsilon\bar{\Gamma}(\theta)}{1+\bar{\Gamma}(\theta)}\right)}\right\rfloor$ *which corresponds to an achievable rate of* $\bar{\Gamma}^*(K(\epsilon,\theta))=\frac{e^{\ln(\epsilon)/K^*(\epsilon,\theta)}-1}{\epsilon-e^{\ln(\epsilon)/K^*(\epsilon,\theta)}}$.

Proof. Please see Appendix A. □

Lemma 3 (Energy constrained cluster size)**.** *For a cluster of energy constrained UEs with* $\bar{\Gamma}(\theta)=\max_i(\bar{\Gamma}_i(\theta))$*, the largest feasible cluster size falls within the range of* $K_{\min}(\epsilon,\theta)\leqslant K(\epsilon,\theta)\leqslant K_{\max}(\epsilon,\theta)$ *where* $K_{\min}(\epsilon,\theta)=\left\lfloor 1+\frac{\ln\left(\frac{\epsilon(1+\bar{\Gamma}(\theta))}{\epsilon-1}\right)-\ln\left(\frac{\bar{\Gamma}(\theta)\sigma^2}{\bar{P}_u g_K}-\frac{1+\epsilon\bar{\Gamma}(\theta)}{1-\epsilon}\right)}{\ln\left(\frac{1+\epsilon\bar{\Gamma}(\theta)}{1+\bar{\Gamma}(\theta)}\right)}\right\rfloor$ *and*

[3] We assume that ϵ cannot be zero in practice. To evaluate the numerical results for $\epsilon\to 0$, we employ the smallest positive normalized floating-point number based on the IEEE Standard for floating-point arithmetic (IEEE 754), i.e., $\epsilon=2.2251\text{e}{-308}$ [34].

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/TCOMM.2019.2927561, IEEE Transactions on Communications

$$K_{\max}(\epsilon,\theta) = \left\lceil 1 + \frac{\ln\left(\frac{\bar{\Gamma}(\theta)\sigma^2}{\bar{P}_u g_1} + \frac{\epsilon(1+\bar{\Gamma}(\theta))}{1-\epsilon}\right) - \ln\left(\frac{1+\epsilon\bar{\Gamma}(\theta)}{1-\epsilon}\right)}{\ln\left(\frac{1+\epsilon\bar{\Gamma}(\theta)}{1+\bar{\Gamma}(\theta)}\right)} \right\rceil.$$

Accordingly, $K^{\star}(\epsilon,\theta) = K_{\min}(\epsilon,\theta)$ is the largest feasible cluster size which is mainly determined by the user with the lowest channel gain.

Proof. Please see Appendix B. □

Once can draw the inference from these lemmas that the largest feasible cluster size increases as $\bar{\Gamma}(\theta)$ and ϵ decreases, that is, NOMA can serve more users with low rates as the SIC efficiency improves. Notice that cluster size analyses in Lemma 1 and Lemma 2 are only valid for UEs with unlimited transmission power as $\boldsymbol{p}^*$ is a solution over the feasible set of $\mathbf{p} > \mathbf{0}$. However, Lemma 3 accounts for power constrained users, where channel gain of the lowest cluster member plays an important role.

B. Cluster Formation Design

Unlike the basic NOMA clusters of size two, one can reap the full benefits of high-spectral efficiency offered by NOMA if the large cluster size is considered. In addition to the enhanced spectral efficiency, increasing the cluster size also reduces the total power consumption of UEs within a BS, that magnifies the efficiency of energy spent per bit. Therefore, our clustering strategy is to exploit the largest feasible cluster size obtained in the previous section. This strategy is especially important to provide the massive connectivity required by the ever-increasing number of devices. When each cluster is allocated to a single RB, for example, basic NOMA clustering can accommodate at most 2Θ UEs at a time. Notice that employing large clusters is eminently suitable for UL-NOMA scheme since UEs do not compete for the BS transmit power as in the DL case. However, a larger cluster size requires more computational power to compute optimum power levels and yields a longer decoding delay as the SIC latency linearly increase with the cluster size [30]. Fortunately, BSs can be equipped with high computational power with more sophisticated receivers with desirable FEF and latency specifications.

Based on the analytical findings in Section III, our strategy on assigning UEs to clusters focuses on maximizing the channel gain disparity among the cluster members to enhance the achievable NOMA gain. Accordingly, algorithmic implementation of these strategies is given in Algorithm 1 where the first line uses Lemma 3 to determine the largest cluster size that is allowable by each UEs within BS_b, i.e., $\kappa_i = \min\left(\bar{K}_b, K_{\min}\right), i \in \mathcal{U}_b$, where $\bar{K}_b$ is a design parameter in order to prevent unnecessarily high delay and computational power due to the large cluster sizes. Accordingly, $(\kappa_i, \forall i)$ values are sorted in the ascending order to generate the vector $\boldsymbol{\kappa} = [\kappa_i |\, i \in \mathcal{U}_b, \kappa_j > \kappa_{j+1}, 1 \leqslant j \leqslant U_b - 1]$. Starting from the larges t cluster size, line 2 increases the number of clusters in BS_b until total number of cluster sizes are no less than $\mathcal{U}_b$, i.e.,

$$C_b = \text{argmin}_{\text{I}} \left\{ \text{I} \left| \sum_{i=1}^{\text{I}} \kappa_i \geqslant U_b \right. \right\} \tag{21}$$

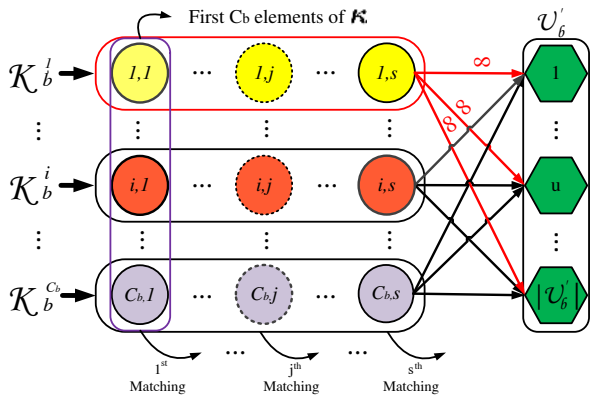


Fig. 2: Illustration of the proposed CF method for sum-rate Maximization.

which provides the least number of clusters and thus the largest size of clusters.

Algorithm 1 Cluster Formation for BS_b, $\forall b$.

Input: $\bar{\boldsymbol{\Gamma}}(\theta)$, $\boldsymbol{h}$
1: $\boldsymbol{\kappa} \leftarrow$ Sort UEs in ascending order as per Lemma 3.
2: $C_b \leftarrow$ Determine the least number of clusters as per (21).
3: $\mathcal{K}_b^i \leftarrow \boldsymbol{\kappa}[i]$, $1 \leqslant i \leqslant C_b$ Predetermination of first cluster members.
4: $\mathcal{U}_b' \leftarrow$ Update the remaining set of UEs.
5: **for** $s = 1 : \left(\left\lceil \frac{U_b}{C_b} \right\rceil - 1\right)$ **do**
6: $\quad \mathcal{E}_i^j(s) \leftarrow$ (22) Calculate edge weights
7: $\quad \mathcal{K}_b^c \leftarrow \min_{\boldsymbol{x}} \sum_{i,j} \mathcal{E}_i^j(s) x_i^j$ (s.t.) $\sum_i x_i^j \leqslant 1, \sum_j x_i^j = 1, x_i^j \in \{0,1\}, i \in [1, C_b], j \in [1, |\mathcal{U}_b'|]$
8: $\quad \mathcal{U}_b' \leftarrow$ Update the remaining set of UEs.
9: **end for**
10: **return** $\mathcal{K}_b^c$, $\forall c$.

As illustrated in Fig. 2, line 3 of Algorithm 1 predetermines i^{th}, $1 \leqslant i \leqslant C_b$, element of $\boldsymbol{\kappa}$ as the first member of i^{th} cluster $\mathcal{K}_b^i$ which has the size of $\boldsymbol{\kappa}_i$. Thereafter, the while loop between lines 6 and 10 iteratively matches clusters with the remaining set of UEs, i.e., $\mathcal{U}_b' = \mathcal{U} - \bigcup_{i=1}^{C_b} \mathcal{K}_b^i$. In line 7 of iteration s, matching weight from i^{th} cluster to j^{th} element of $\mathcal{U}_b'$ is calculated as

$$\mathcal{E}_i^j(s) = \begin{cases} \frac{\inf\left(\widehat{\mathcal{H}}_i^j\right)}{h_j^b} + \frac{h_j^b}{\sup\left(\check{\mathcal{H}}_i^j\right)}, i \in \mathcal{K}_b, j \in \mathcal{U}_b' & \text{, if } \boldsymbol{\kappa}_i > s \\ \infty & \text{, otherwise} \end{cases} \tag{22}$$

where $\widehat{\mathcal{H}}_i^j = \{h_k^b | h_k^b \geqslant h_j^b, k \in \mathcal{K}_b^i, j \in \mathcal{U}'\}$ and $\check{\mathcal{H}}_i^j = \{h_k^b | h_k^b \leqslant h_j^b, k \in \mathcal{K}_b^i, j \in \mathcal{U}'\}$ are set of cluster members with higher and lower channel gain than the $\text{UE}_j \in \mathcal{U}_b'$, respectively[4]. The first and second term of (22) favors for new members who give a desirable channel gain disparity between $\text{UE}_j \in \mathcal{U}_b'$ and current cluster members with high and low channel gains. Notice that clusters that reached to its maximum affordable size are taken out of consideration by setting their edge weights to infinity. Line 8 executes maximum weighted bipartite matching, $\mathcal{K}_b^c \leftarrow \min_{\boldsymbol{x}} \sum_{i,j} \mathcal{E}_i^j(s) x_i^j$ (s.t.) $\sum_i x_i^j \leqslant$

[4]Notice that either $\widehat{\mathcal{H}}_i^j = \varnothing$ or $\check{\mathcal{H}}_i^j = \varnothing$ happens for $s = 1$. Since the order theory of the real analysis tells us that $\inf(\varnothing) = \infty$ and $\inf(\varnothing) = -\infty$, we ignore the first (second) term of (22) if $\widehat{\mathcal{H}}_i^j = \varnothing$ ($\check{\mathcal{H}}_i^j = \varnothing$) occurs.

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/TCOMM.2019.2927561, IEEE Transactions on Communications

$1, \sum_j x_i^j = 1, x_i^j \in \{0,1\}, i \in [1, C_b], j \in [1, |\mathcal{U}_b'|]$, which is in the form of a rectangular assignment problem and can be solved in cubic order. Algorithm 1 is run by each BS independent from others and its overall complexity can be given as $\mathrm{O}\left(U_b \log U_b + \sum_{s=1}^{\left\lceil \frac{U_b}{C_b} \right\rceil - 1} (U_b - sC_b)^3\right)$ where the first and second terms are due to sorting and matching operations in lines 1 and 8, respectively. Since the second term is more dominant, the proposed clustering solutions has cubic time complexity. On the other hand, exhaustively checking all clustering sizes and corresponding user combination $\sum_{k=2}^{Ub} \binom{U_b}{k} \approx 2^{U_b}$ which yields an exponential time complexity.

VI. α-Fair Resource Allocation

In this section, we handle the RA problem by decoupling it into two stages: In the former, a slave problem is defined for each cluster such that optimal power allocations are obtained in closed-form for given cluster formations and bandwidths. In the latter, each slave problem reports its obtained utility which is exploited by a master problem to update cluster bandwidths.

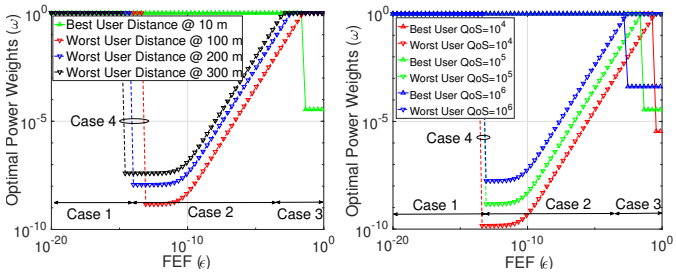
A. Slave Problems: Power Allocation

Power allocation problem of clusters can be formulated as in (VI-A) where we omit BS and cluster indices for the sake of simplicity without loss of generality.

$$\mathbf{S}: \max_{\boldsymbol{\omega}} \quad \sum_{i=1}^{K} \pi_i(\boldsymbol{\omega})$$
$$\mathrm{S}_\mathrm{i}^1: \quad \text{s.t.} \quad \omega_i \leqslant 1, \ \forall i$$
$$\mathrm{S}_\mathrm{i}^2: \qquad 0 \leqslant \omega_i h_i - \bar{\Gamma}_i \left(\epsilon \sum_{j=1}^{i-1} \omega_j h_j + \sum_{k=i+1}^{K} \omega_k h_k + \varrho \right), \forall i$$

which can be locally solved by each cluster member for given cluster bandwidth and channel gains of other cluster members. In order to derive the closed-form expressions for optimal power allocations, we first apply dual decomposition method to the slave problems. Accordingly, Lagrangian function of $\mathbf{S}$ is given in (23) where λ_i and $\mu_i, \forall i$, are Lagrange multipliers. Taking derivatives of Lagrangian function with respect to ω_i, λ_i and μ_i, Karush-Kuhn-Tucker (KKT) conditions can be obtained as in (24)-(25).

KKT conditions are first-order necessary conditions for a nonlinear programming solution to be optimal, which is still subject to satisfaction of some regularity conditions. In particular, if all equality and inequality constraints are affine functions, i.e., linearity constraint qualification is held, no other regularity condition is needed. This is indeed the case for $\mathbf{S}$ as all constraints are affine functions of $\boldsymbol{\omega}$. In the slave problem, there exists a total of $2K$ Lagrange multipliers that can be categorized into two subsets $\mathcal{S}_1 = \{\lambda_i | 1 \leqslant i \leqslant K\}$ and $\mathcal{S}_2 = \{\mu_i | 1 \leqslant i \leqslant K\}$. Therefore, each slave problem requires the KKT condition verification of 2^{2K} Lagrange multiplier combinations. Even though this is computationally impractical, we fortunately need to check only 2^K combinations [12], [36] for the following reasons: Notice



(a) (b)

Fig. 3: Optimal power allocations of a basic cluster vs. FEF levels; a) different channel gain disparity and b) QoS constraint scenarios.

that each UE would transmit at the maximum transmission power in case of no interference, i.e, OMA. However, optimal power levels of NOMA can either be determined by CSCs or maximum transmission power according to SINR requirements and achievable capacity of UEs. That is, UE_i can be active either on maximum transmission power or CSCs at the optimal point. Hence, we need to consider the following solution set $\mathcal{S} = \{\lambda_i \text{ or } \mu_i | i \in [1, K]\}$ in order to obtain a closed-form solution. For a basic NOMA cluster, combinations of solution set can be given as $\{\lambda_1, \lambda_2\}$, $\{\lambda_1, \mu_2\}$, $\{\mu_1, \lambda_2\}$, and $\{\mu_1, \mu_2\}$. Furthermore, $\mathcal{S}_\lambda = \mathcal{S} - \mathcal{S}_2$ and $\mathcal{S}_\mu = \mathcal{S} - \mathcal{S}_1$ represents the subset of the solution set $\mathcal{S}$ which define cluster members active at λ and μ, respectively. Finally, $\mathcal{I}_\lambda$ and $\mathcal{I}_\mu$ denotes the index set of $\mathcal{S}_\lambda$ and $\mathcal{S}_\mu$, respectively. For example, for the solution set of $\mathcal{S} = \{\mu_1, \lambda_2, \lambda_3, \mu_4, \lambda_5, \mu_6\}$, we have $\mathcal{S}_\lambda = \{\lambda_2, \lambda_3, \lambda_5\}$, $\mathcal{S}_\mu = \{\mu_1, \mu_4, \mu_6\}$, $\mathcal{I}_\lambda = \{2, 3, 5\}$ and $\mathcal{I}_\mu = \{1, 4, 6\}$.

For example, let us consider $\mathcal{S} = \{\lambda_1, \mu_2, \lambda_3, \mu_4\}$, then the power allocations can be derived from active primal constraints, i.e., $\{\mathrm{S}_1^1, \mathrm{S}_2^2, \mathrm{S}_3^1, \mathrm{S}_4^2\}$, which also requires the satisfaction of corresponding primal KKT conditions, i.e., $\{\mathrm{S}_1^2, \mathrm{S}_2^1, \mathrm{S}_3^2, \mathrm{S}_4^1\}$. That is, active primal constraints form the KKT conditions while inactive constraints are used for calculating the corresponding power allocations. Accordingly, we tabulate power allocations and corresponding KKT conditions for cluster sizes 2 and 3 in Table I where the first column indicates the cluster size K, the second column presents 2^K solution set cases and corresponding necessary conditions, and finally the last row provides the closed-form optimal power allocations. Excluding the first case, both power allocations and necessary conditions are functions of three parameters; ϵ, $\bar{\Gamma}$, and channel gain disparity. Based on these parameters, while some cases can be infeasible due to the violation of the necessary conditions, there might me multiple cases which satisfy the constraint with different performance.

Before we explain how the optimal case is determined, we consider an exemplary basic NOMA cluster size of two in order to have a deeper insight into how the power allocation strategy changes with different parameters which have a direct impact on the constraints, i.e., necessary conditions. The optimal power weights versus different FEF levels under various channel gain disparity and QoS constraints are shown in Fig. 3a and Fig. 3b, respectively. As it is already tabulated in Table 1, there exist four cases: In case 1, both UEs transmit

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/TCOMM.2019.2927561, IEEE Transactions on Communications

$$L(\boldsymbol{\omega},\boldsymbol{\lambda},\boldsymbol{\mu}) = \frac{1}{1-\alpha}\sum_{i=1}^{K}\left(R_i^{1-\alpha}(\boldsymbol{\omega})-1\right)+\sum_{i=1}^{K}\lambda_i(1-\omega_i)+\sum_{i=1}^{K}\mu_i\left(\omega_i g_i-\bar{\Gamma}_i\left(\epsilon_i\sum_{j=1}^{i-1}\omega_j g_j-\sum_{k=i+1}^{K}\omega_k g_k-\varrho\right)\right) \tag{23}$$

$$\frac{\partial\mathcal{L}}{\partial\omega_i^\star} = W\theta\left\{\frac{R_i^{-\alpha}(\boldsymbol{\omega})}{\sum_{k=1}^{i-1}\epsilon h_k\omega_k+\sum_{l=i}^{K}\omega_l h_l+\varrho}-\sum_{j=1}^{i-1}\frac{\omega_j h_j R_j^{-\alpha}(\boldsymbol{\omega})}{\left(\sum_{k=1}^{j-1}\epsilon h_k\omega_k+\sum_{l=j}^{K}\omega_l h_l+\varrho\right)\left(\sum_{k=1}^{j-1}\epsilon h_k\omega_k+\sum_{l=j}^{K}\omega_l h_l+\varrho\right)}\right.$$
$$\left.-\sum_{j=i+1}^{K}\frac{\epsilon\omega_j h_j R_j^{-\alpha}(\boldsymbol{\omega})}{\left(\sum_{k=1}^{j-1}\epsilon h_k\omega_k+\sum_{l=j}^{K}\omega_l h_l+\varrho\right)\left(\sum_{k=1}^{j-1}\epsilon h_k\omega_k+\sum_{l=j}^{K}\omega_l h_l+\varrho\right)}\right\}$$
$$-\lambda_i+(1-\Gamma_i)\mu_i-\sum_{j=1}^{i-1}\Gamma_j\mu_j-\epsilon\sum_{j=i+1}^{\mathrm{K}}\Gamma_j\mu_j\geqslant 0,\forall i, \tag{24}$$

$$\frac{\partial\mathcal{L}}{\partial\lambda_i^\star} = 1-\omega_i\geqslant 0,\ \text{if}\ \lambda_i^\star\geqslant 0,\frac{\partial\mathcal{L}}{\partial\mu_i^\star} = \omega_i-\left(\sum_{j=1}^{i-1}\omega_j+\epsilon_i\sum_{k=i+1}^{K_c^r}\omega_k+\rho_i\right)(q_i-1)\geqslant 0,\ \text{if}\ \mu_i^\star\geqslant 0. \tag{25}$$

TABLE I: Necessary conditions and closed-form power allocations (Please see Appendix C for the proof).

K	Necessary Conditions	Power Allocations
2	$\mathcal{S}=\{\lambda_1,\lambda_2\}:\mathrm{S}_l^2,\mu_l>0,l=1,2.$	$\omega_1=\omega_2=1$
	$\mathcal{S}=\{\lambda_1,\mu_2\}:\mathrm{S}_{\mathrm{k}}^1,\lambda_k>0,k=2.\mid\mathrm{S}_l^2,\mu_l>0,l=1.$	$\omega_1=1 \mid \omega_2=\frac{\bar{\Gamma}_2(h_1\epsilon+\varrho)}{h_2}$
	$\mathcal{S}=\{\mu_1,\lambda_2\}:\mathrm{S}_{\mathrm{k}}^1,\lambda_k>0,k=1.\mid\mathrm{S}_l^2,\mu_l>0,l=2$	$\omega_1=\frac{\bar{\Gamma}_1(h_2+\varrho)}{h_1} \mid \omega_2=1$
	$\mathcal{S}=\{\mu_1,\mu_2\}:\mathrm{S}_k^1,\lambda_k>0,k=1,2.$	$\omega_1=\frac{\varrho}{\frac{h_1}{\bar{\Gamma}_1}-a_1h_2} \mid \omega_2=\frac{a_1\varrho}{\frac{h_1}{\bar{\Gamma}_1}-a_1h_2}$
3	$\mathcal{S}=\{\lambda_1,\lambda_2,\lambda_3\}:\mathrm{S}_l^2,\mu_l>0,l=1,2,3.$	$\omega_1=\omega_2=\omega_3=1$
	$\mathcal{S}=\{\lambda_1,\lambda_2,\mu_3\}:\mathrm{S}_{\mathrm{k}}^1,\lambda_k>0,k=3.\mid\mathrm{S}_l^2,\mu_l>0,l=1,2.$	$\omega_1=\omega_2=1 \mid \omega_3=\frac{\bar{\Gamma}_3(h_1\epsilon+h_2\epsilon+\varrho)}{h_3}$
	$\mathcal{S}=\{\lambda_1,\mu_2,\lambda_3\}:\mathrm{S}_{\mathrm{k}}^1,\lambda_k>0,k=2.\mid\mathrm{S}_l^2,\mu_l>0,l=1,3.$	$\omega_1=\omega_3=1 \mid \omega_2=\frac{\bar{\Gamma}_2(h_1\epsilon+h_3+\varrho)}{h_2}$
	$\mathcal{S}=\{\lambda_1,\mu_2,\mu_3\}:\mathrm{S}_{\mathrm{k}}^1,\lambda_k>0,k=2,3.\mid\mathrm{S}_l^2,\mu_l>0,l=1.$	$\omega_1=1 \mid \omega_2=\frac{h_1\epsilon+\varrho}{\frac{h_2}{\bar{\Gamma}_2}-a_1h_3} \mid \omega_3=\frac{a_1(h_1\epsilon+\varrho)}{\frac{h_2}{\bar{\Gamma}_2}-a_1h_3}$
	$\mathcal{S}=\{\mu_1,\lambda_2,\lambda_3\}:\mathrm{S}_{\mathrm{k}}^1,\lambda_k>0,k=1.\mid\mathrm{S}_l^2,\mu_l>0,l=2,3.$	$\omega_1=\frac{\bar{\Gamma}_1(h_2+h_3+\varrho)}{h_1} \mid \omega_2=\omega_3=1$
	$\mathcal{S}=\{\mu_1,\lambda_2,\mu_3\}:\mathrm{S}_{\mathrm{k}}^1,\lambda_k>0,k=1,3.\mid\mathrm{S}_l^2,\mu_l>0,l=2.$	$\omega_1=\frac{c_2h_3+h_2+\varrho}{\frac{h_1}{\bar{\Gamma}_1}-a_1h_3} \mid \omega_2=1 \mid \omega_3=\frac{a_1(c_2h_3+h_2+\varrho)}{\frac{h_1}{\bar{\Gamma}_1}-a_1h_3}+c_2$
	$\mathcal{S}=\{\mu_1,\mu_2,\lambda_3\}:\mathrm{S}_{\mathrm{k}}^1,\lambda_k>0,k=1,2.\mid\mathrm{S}_l^2,\mu_l>0,l=3.$	$\omega_1=\frac{h_3+\varrho}{\frac{h_1}{\bar{\Gamma}_1}-a_1h_2} \mid \omega_2=\frac{a_1(h_3+\varrho)}{\frac{h_1}{\bar{\Gamma}_1}-a_1h_2} \mid \omega_3=1$
	$\mathcal{S}=\{\mu_1,\mu_2,\mu_3\}:\mathrm{S}_{\mathrm{k}}^1,\lambda_k>0,k=1,2,3.$	$\omega_1=\frac{\varrho}{\frac{h_1}{\bar{\Gamma}_1}-a_1h_2-a_1a_2h_3} \mid \omega_2=\frac{a_1\varrho}{\frac{h_1}{\bar{\Gamma}_1}-a_1h_2-a_1a_2h_3}$ $\omega_3=\frac{a_1a_2\varrho}{\frac{h_1}{\bar{\Gamma}_1}-a_1h_2-a_1a_2h_3}$

at maximum power. In case 2, the worst UE is active at the QoS constraint whereas the best user keeps transmitting at the maximum power, which is in the opposite direction for case 3. In the last case, both UEs have power levels that exactly and barely satisfy their QoS demands. As it is obvious in Fig. 3, power levels and case regions vary with FEF levels, channel gain disparity, and QoS demands. While we observe case 1 and case 3 in very low and very high FEF levels, respectively, intermediate FEF levels operate on case 2. On the other hand, case 4 is observed during the interval where optimal case is in transition from case 1 to case 2. Notice that channel gain disparity and QoS constraints have significant impacts on both optimal power levels and the points where cases start and end.

At this point, let us explain how Table I can be used to decide on the optimal case: First, power allocations of each case are computed by expressions on the rightmost column, and then substituted into the corresponding constraints in the central column to verify if the corresponding KKT conditions are satisfied. Thereafter, optimal power allocations are determined by the case which gives the highest objective value among the cases who satisfy the KKT conditions. Therefore, the worst case complexity is given as $\mathrm{O}(2^K+K\log K)$ where

Lemma 4. *Given that necessary conditions are satisfied, closed-form power allocations of cluster members is given as*

$$\omega_i = \begin{cases} 1 & ,for\ \forall i \in \mathcal{I}_\lambda, \\ \left(\prod_{\substack{1 \leqslant j < i \\ j \in \mathcal{I}_\mu}} a_j\right) \omega_{\text{mind}} + \sum_{\substack{1 \leqslant j < i \\ j \in \mathcal{I}_\mu}} \left(\prod_{\substack{j < k < i \\ k \in \mathcal{I}_\mu}} a_k\right) b_j & ,for\ \forall i \in \mathcal{I}_\mu, i > \text{mind}. \end{cases}, \tag{26}$$

where $\omega_{\text{mind}} = \dfrac{c_{\text{mind}} + \sum_{\substack{\text{mind} < k \leqslant K \\ k \in \mathcal{I}_\mu}} h_k \left(\sum_{\substack{1 \leqslant j < k \\ j \in \mathcal{I}_\mu}} \left(\prod_{\substack{l > j \wedge j < k \\ l \in \mathcal{I}_\mu}} a_l \right) c_j \right)}{\frac{h_{\text{mind}}}{\Gamma_{\text{mind}}} - \sum_{\substack{\text{mind} < k \leqslant K \\ k \in \mathcal{I}_\mu}} h_k \left(\prod_{\substack{1 \leqslant j < k \\ j \in I_\mu}} a_j \right)}$, $\text{mind} \triangleq \text{argmin}(\mathcal{I}_\mu)$ *is the minimum index of UEs within* $\mathcal{S}_\mu$, $c_{\text{mind}} = \epsilon \sum_{\substack{1 \leqslant j < \text{mind} \\ j \in \mathcal{I}_\lambda}} h_j + \sum_{\substack{\text{mind} < j \leqslant K \\ j \in \mathcal{I}_\lambda}} h_j + \sum_{\substack{\text{mind} < j \leqslant K \\ j \in \mathcal{I}_\mu}} \omega_j h_j$, $a_i = \dfrac{h_{\max_i} \left(\epsilon + \frac{1}{\Gamma_{\max_i}} \right)}{h_i \left(1 + \frac{1}{\Gamma_i} \right)}$, $b_i = \dfrac{(\epsilon - 1)}{h_i \left(1 + \frac{1}{\Gamma_i} \right)} \sum_{\substack{\max_i < j < i \\ j \in \mathcal{I}_\lambda}} h_j$, *and* $\max_i = \text{argmax}\{m | m \in \mathcal{I}_\mu, m < i\}$ *is the maximum index of* $\mathcal{S}_\mu$ *among the indices less than* i.

Proof. Please see Appendix C. □

the first term is the cost of calculating and checking 2^K cases and the second term is the cost of sorting and selecting the best case. Since the complexity of the first term dominates that of the second, overall complexity can be approximated by $\text{O}(2^K)$. Generalizing Table I, Lemma 4 provides the closed-form expression for optimal power allocations for an imperfect NOMA cluster of size K.

B. Master Problem: Bandwidth Allocation

Following the optimal power allocation of the slave problems, BSs report the achieved SINR levels of cluster members to the MBS which then updates the bandwidth allocations as follows

$$\begin{aligned} \mathbf{M}: \max_{\boldsymbol{\theta}} \quad & \frac{1}{1-\alpha} \sum_{\forall (b,c,i)} \left[\left(\theta_b^c \vartheta_{b,c}^i \right)^{1-\alpha} - 1 \right] \\ \text{M}_\text{i}^1: \quad \text{s.t.} \quad & \sum_{b,c} \theta_{b,c} \leqslant \Theta, \ \forall i \\ \text{M}_\text{i}^2: \quad & \bar{R}_i \leqslant \theta_b^c \vartheta_{b,c}^i, \ \forall i \in \mathcal{K}_b^c, \forall b, \forall c. \end{aligned}$$

where $\vartheta_{b,c}^i \triangleq W \log_2(1 + \Gamma_{b,c}^i)$ is the achieved utility of clusters and given by the slave problems. An effective method of solving this problem is unintegerizing the integer valued optimization variable θ_b^c. In this manner, $\mathbf{M}$ reduces to a convex optimization problem and fractional part of the optimal bandwidth allocations can be handled by RB scheduling mechanisms.

Proposed distributed α-fair resource allocation framework is summarized in Algorithm 2 which is indeed a detailed algorithmic version of Fig. 1. In Algorithm 2, BSs are only required to know channel gains of their own UEs. Following the initialization of the cluster bandwidths in line 2, the while loop between lines 3 and 11 iteratively forms clusters, obtains power allocations and update bandwidths until a termination term is not reached. In line 4, each BS first forms its clusters based on the steps given in Algorithm 1. According to the CF outcome, BSs solve slave problems to calculate the optimal power levels as explained in the previous section in lines 5 and 6, then transmits optimal power allocations to UEs in line 7. Thereafter, BSs share observed utilities with the MBS in line 8, which is followed by a bandwidth allocation update and dissemination in line 9.

Algorithm 2 Distributed α-Fair Resource Allocation

Input: Channel gains
1: $t \leftarrow 0$
2: $\boldsymbol{\theta}(\boldsymbol{k}) \leftarrow$ Initialize the bandwidth allocations, $\forall b, c$.
3: **while** $t \in \mathcal{T}$ **do**
4: $\quad \boldsymbol{\delta_b} \leftarrow \text{BS}_b$ forms its clusters based on Algorithm 1.
5: $\quad \boldsymbol{\omega_b^c(t)} \leftarrow$ Check power levels & conditions.
6: $\quad \boldsymbol{\omega_{b,c}^\star(t)} \leftarrow$ Select the best cases for optimal power allocation.
7: $\quad \text{UE}_i \leftarrow \boldsymbol{\omega_{b,c}^i}$; UE_i receives its power level from $\text{BS}_b, \forall i \in \mathcal{U}_b$.
8: $\quad \text{BS}_0 \leftarrow \boldsymbol{\vartheta_b^c}$; The MBS receives the utilities from $\text{BS}_b, \forall b$
9: $\quad \text{BS}_b \leftarrow \boldsymbol{\theta(t+1)}$; The MBS updates and disseminates bandwidths to $\text{BS}_b, \forall b$.
10: $\quad t \leftarrow t + 1$
11: **end while**
12: **return** Power and bandwidth allocations

It is necessary to point out that the first step of next iteration starts with reclustering if there is a change in cluster size or a significant variation in channel gains[5]. Since all steps between lines 4 and 8 are executed by BSs in a parallel fashion[6], the computational complexity for each BS is mainly driven by clustering and power control steps whose time complexity are given in Section V-B and Section VI-A, respectively. Although the MBS has an extra duty for solving the master problem M, complexity of solving a convex problem is negligible in comparison with clustering and power allocation.

Notice that there are two types of message passing in Algorithm 2: The former occurs between BS_b and its users to share optimal power allocations, which is in the order of U_b. The latter takes place between smallcells and the MBS to receive bandwidth updates and report obtained utilities (line 9), which is in order of the total number of clusters, C_b. Therefore, proposed distributed method has a low communication overhead.

[5]While channel gain variations can be caused by user mobility, cluster size varies either with bandwidth or QoS updates.

[6] If a BS fails to implement the proposed scheme, it can switch to OMA scheme until it recovers from the failure.

TABLE II: Table of Parameters

Par.	Value	Par.	Value	Par.	Value
η_b^u	3.76	ϵ	10^{-7}	β	0.025
N_0	$-174\ dBm$	K_b	10	P_u	$23\ dBm$
W	180 kHz	U	100	P_s	$30\ dBm$
Θ	100	B	10	P_m	$46\ dBm$

Algorithm 2 starts with clustering and then proceed with the power allocation. Even though reversing this order can be thought as an alternative method, it is challenging due to several practical reasons: First, initial cluster bandwidths are necessary to calculate the feasible cluster sizes, that is the first step of clustering. Second, initial cluster bandwidths are also necessary for the power allocation problem because QoS constraints depends on the available cluster bandwidth. Since it is challenging to solve these two main subproblems without an initial bandwidth allocation at the first iteration, we follow the former approach.

VII. Numerical Results and Analysis

For the simulations, we consider U UEs and B SBSs uniformly distributed over a cell area of $500\ m \times 500\ m$ MBS. QoS requirements of UEs are randomly determined with a mean of 1 Mbps. All results are obtained by averaging over 200 network scenarios. The composite channel gain, h_b^i, between BS$_b$ and UE$_i$ is given as

$$h_b^i = A_b^i \delta_{b,i}^{-\eta_b^i} 10^{\xi_b^i/10} \mathbb{E}\{|g_b^i|^2\} \tag{27}$$

where A_b^i is a constant related to antenna parameters, $\delta_{b,i}$ is the distance between the nodes, η_b^i is the path loss exponent, $10^{\xi_b^i/10}$ represents the log-normally distributed shadowing, ξ_b^i is a normal random variable representing the variation in received power with a variance of ς_b^i, i.e., $\xi_b^i \sim \mathcal{N}(0, \varsigma_b^i)$, $\tilde{h}_b^i$ is the complex channel fading coefficient, $\mathbb{E}\{\cdot\}$ is the expectation to average small scale fading out, and $\mathbb{E}\{|g_b^i|^2\}$ is assumed to be unity. Unless it is stated explicitly otherwise, we use the default simulation parameters given in Table II.

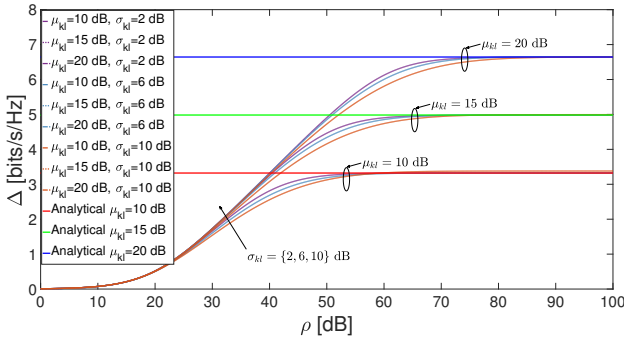


Fig. 4: Impact of channel gain disparity on NOMA gain.

A. Impacts of Channel Gain Disparity and Decoding Order

Fig. 4 compares the analytical findings obtained in Section III with the simulations where the reference user, UE$_k$, placed 10 m away from the MBS with a deviation of 2 dB shadow fading. UE$_\ell$ is placed in (100, 300, 1000) m away with (0, 4, 8) dB deviation which results in $\mu_{kl} = \{10, 15, 20\}$ dB and $\sigma_{kl} = \{2, 6, 10\}$ dB, respectively. As $\rho = \bar{P}_u/N_0 B$ reaches up to 100 dB, simulation converges to the upper bound in (14). Please note that for P_u and N_0 given in Table II, practical values of ρ ranges from 245 dB to 320 dB for bandwidths ranging from 1 Hz to 20 MHz. That is, the analytical upper bound is tight enough for practical values of ρ.

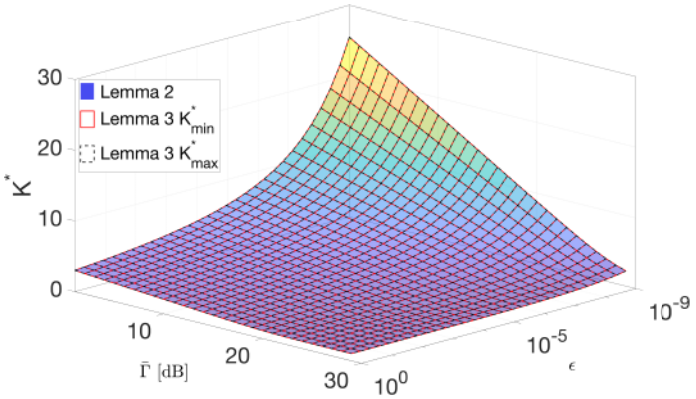


Fig. 5: The largest feasible cluster size vs. different FEF and CSCs values.

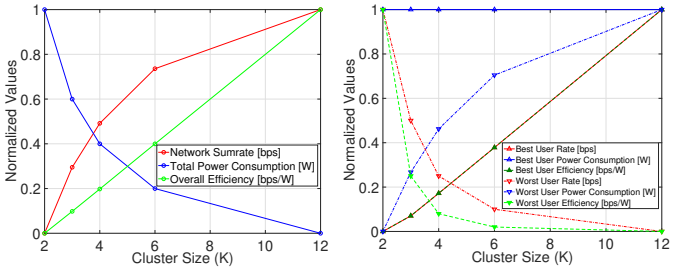
B. The Largest Feasible Cluster Size Analysis

Fig. 5 shows the maximum feasible cluster size that can be handled by a single RB with respect to different $\bar{\Gamma}$ and ϵ values, where we do not use ceiling and floor functions in the lemmas for a better comparison. It is obvious that the cluster size increases as $\bar{\Gamma}$ and ϵ decrease, that is, NOMA can serve more users with low rates as the SIC efficiency improves. As a numerical example, a single RB with $\epsilon = 10^{-5}$ can serve 3 and 4 UEs each with 0.5 Mbps and 1 Mbps, respectively. Fig. 5 also compares the energy constrained cluster size with the unconstrained cluster size, where the weakest UE is located at the cell-edge. From numerical results, we observe that the cluster size difference between the two cases is negligible for practical channel gain values. Hence, changes in the largest cluster size are primarily affected by changes in cluster bandwidth and/or QoS demands.

C. Spectral and Energy Efficiency

To investigate the impacts of cluster size on spectral and power efficiency, let us consider a single BS with 12 UEs which can be grouped into $\{6, 4, 3, 2, 1\}$ clusters with corresponding sizes of $\{2, 3, 4, 6, 12\}$. The normalized values for spectral efficiency, total power consumption, and energy efficiency are shown in Fig. 6 where normalization is done for each curve individually (each has different units) using feature scaling, i.e., $x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$, where x' is the normalized value, x is the actual value before the normalization, and $x_{\min}$ ($x_{\max}$) is the minimum (maximum) of all values of the curve before the normalization. The available bandwidth is uniformly divided between the cluster, thus, sumrate and spectral efficiency are both illustrated with red colored curve. Notice in Fig. 6 that curves are not comparable to each other

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/TCOMM.2019.2927561, IEEE Transactions on Communications



(a) (b)

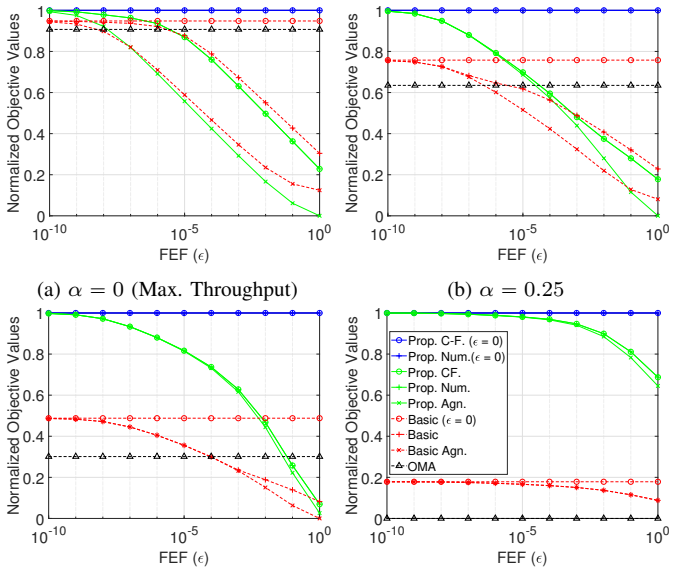
Fig. 6: Impacts of cluster size on spectrum and energy efficiency.

as a point in a curve is relative to other points in the same curve. Throughput the paper we plot figures with normalized values for two reasons; to reduce the number of figures by displaying different curves in different units and to provide a clear comparison in a 0-1 scale is intuitive to infer changes in percentage.

In Fig. 6a, increasing the cluster size obviously enhances the overall spectral efficiency while it has a diminishing impact on the total power consumption of the clusters. For example, the number of UEs transmitting at the maximum power (i.e., the best UE) for cluster sizes of 2 and 3 is 6 and 4 (i.e., the number of clusters), respectively. As a result, cluster size 3 requires %40 less power consumption while providing %30 more sumrate than the basic NOMA, that yields a higher energy efficiency in units of bps/W. Let us now focus on Fig. 6b where we depict the individual performance metrics of the best and worst UEs. While the best UEs keep transmitting at the same power level, their rates and efficiency increase with the cluster size since the bandwidth increases with decreasing number of clusters. However, this behavior follows an opposite direction for the worst UE case. Although the proposed solution allocates powers and bandwidths by taking the QoS constraints of all UEs into consideration, decreasing trend of the worst UE's data rate may cause coverage issues for large cluster sizes in large cells. In particular, providing the demanded QoS for cell-edge macro cell users may hinders the service coverage. At this point, decoupling the DL and UL user association can help in a great extend, which is already investigated and explained in Fig. 9. It is important to shed lights on the tradeoff between the worst and best case user performances. The worst case performance can be enhanced by setting a higher QoS requirement which naturally decreases the achievable rate of the best UE and thus the cluster sumrate. Nonetheless, this could yield a lower best case UE spectral efficiency than that is achievable by OMA, i.e., individual operation of the best case UE. Therefore, a good compromise must be forged to incentivize both UEs to enhance overall spectral efficiency of the network by using NOMA. To this end, cellular network operators can settle certain marketing policies to outline the rules for QoS setting which is satisfactory for both UEs.

D. Impacts of Network Parameters on the NOMA Performance

For the sake of a better comparison, let us consider the following cases: 1) *Prop. CF ($\epsilon = 0$)*: Proposed CF algorithm



(a) $\alpha = 0$ (Max. Throughput) (b) $\alpha = 0.25$

(c) $\alpha = 0.5$ (d) $\alpha \to 1$ (Prop. Fairness)

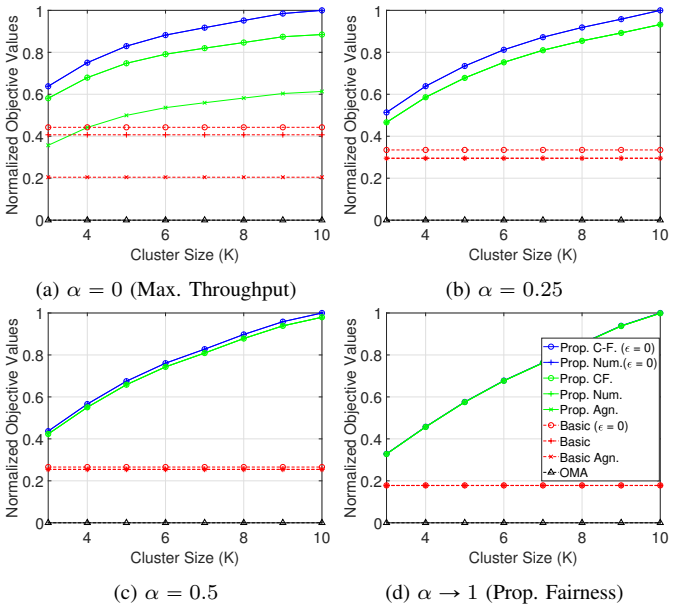
Fig. 7: Normalized network sum-rate vs. FEF levels ϵ.

under perfect NOMA scheme which is obtained by the closed-form expression given in Lemma 4 and drawn by blue colored —○—, 2) *Prop. Num. ($\epsilon = 0$)*: This case is to check the validity of the previous case and drawn by blue colored —+—, 3) *Prop. CF*: Proposed CF algorithm under imperfect NOMA scheme which is obtained by the closed-form expression given in Lemma 4 and drawn by green colored —○—, 4) *Prop. Num.*: This case is to check the validity of the previous case and drawn by green colored —+—, 5) *Prop. Agn.* : The agnostic case is used to show the consequences of treating an imperfect NOMA as perfect by falsely assuming $\epsilon = 0$ and drawn by green colored —×—, 6) *Basic ($\epsilon = 0$)/Basic/Basic Agn.*: This case compares the basic NOMA cluster of size two with the proposed case in 3/4/5 and drawn by red colored —○—/—+—/—×—, and 7) *OMA* corresponds to traditional OMA scheme where entire bandwidth is equally shared among the users and drawn by black colored —△—.

We demonstrate normalized network performance with respect to different network parameters in Fig. 7 - Fig. 11 where normalized objective value is obtained via feature scaling, i.e., $x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$, where x' is the normalized value, x is the value before the normalization, and $x_{\min}$ ($x_{\max}$) is the minimum (maximum) of all values within the figure before the normalization [7] It is common for all subfigures that the proposed solution provides a superior performance in comparison with the traditional basic NOMA and OMA schemes in all cases. This is mainly because of allowing a large number of cluster size, which enhances the spectral efficiency of the network. On the other hand, the basic NOMA scheme delivers a performance in between the proposed solution and OMA scheme. Noting that obtained closed-form expressions perfectly match with numerical solutions, the agnostic approach deteriorates the network performance, which goes even

[7] Although Fig. 7 - Fig. 11 show the network sumrate as a product of the optimized bandwidth and spectral efficiency, readers can also have an insight into the spectral efficiency trends under different network settings.

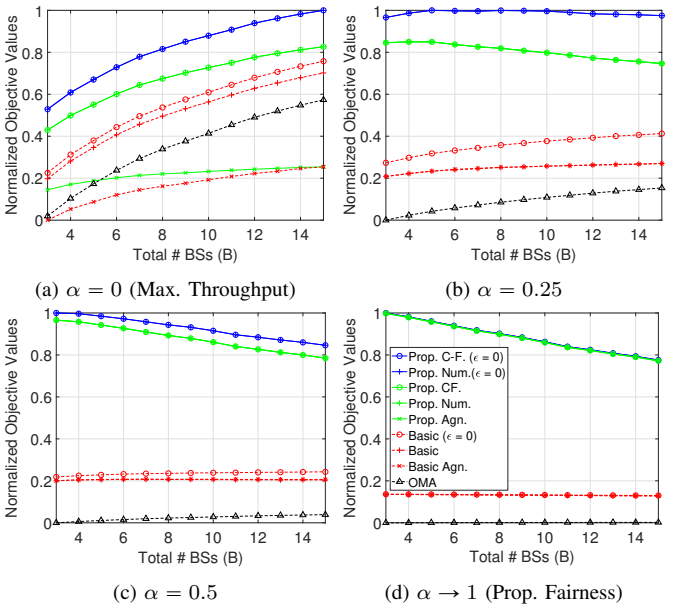
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2019.2927561, IEEE Transactions on Communications



(a) $\alpha = 0$ (Max. Throughput) (b) $\alpha = 0.25$

(c) $\alpha = 0.5$ (d) $\alpha \to 1$ (Prop. Fairness)

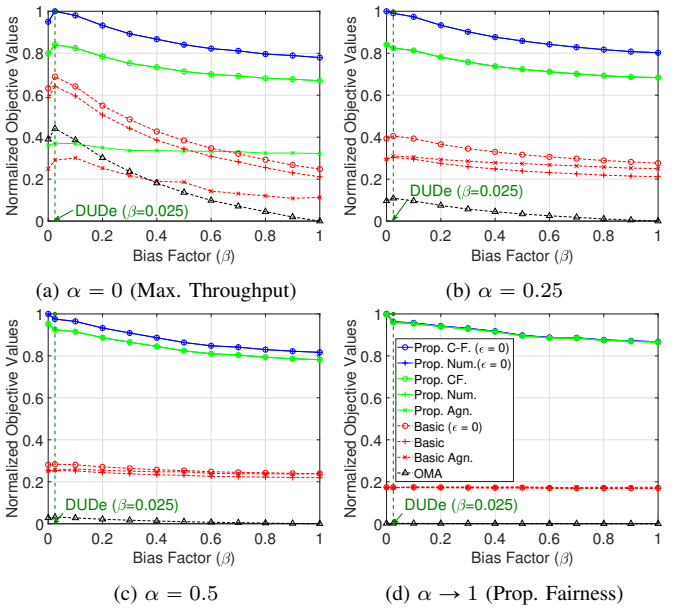
Fig. 8: Normalized network sum-rate vs. affordable cluster size $\bar{K}$.



(a) $\alpha = 0$ (Max. Throughput) (b) $\alpha = 0.25$

(c) $\alpha = 0.5$ (d) $\alpha \to 1$ (Prop. Fairness)

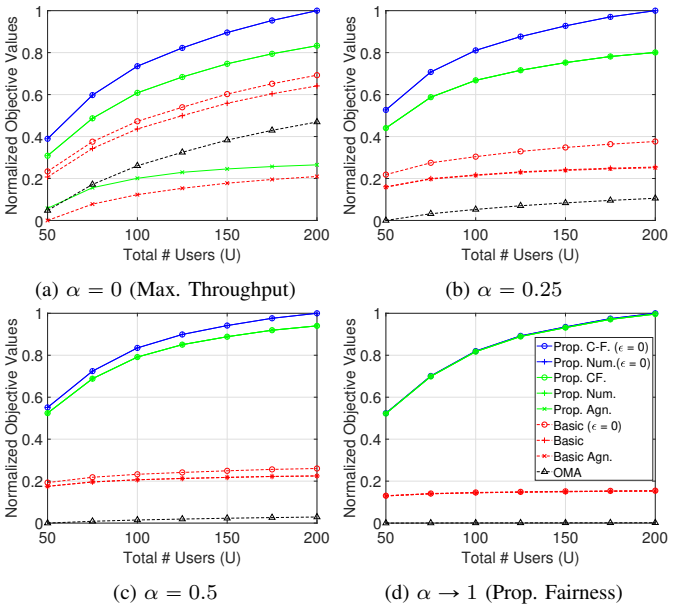
Fig. 10: Normalized network sum-rate vs. total number of SBSs B



(a) $\alpha = 0$ (Max. Throughput) (b) $\alpha = 0.25$

(c) $\alpha = 0.5$ (d) $\alpha \to 1$ (Prop. Fairness)

Fig. 9: Normalized network sum-rate vs. Bias factor β



(a) $\alpha = 0$ (Max. Throughput) (b) $\alpha = 0.25$

(c) $\alpha = 0.5$ (d) $\alpha \to 1$ (Prop. Fairness)

Fig. 11: Normalized network sum-rate vs. total number of UEs U.

below the OMA scheme in certain cases, especially in the maximum throughput case.

Let us start our investigation with the influence of FEF levels on the network performance under different α scenarios as shown in Fig. 7. We involve ourselves in FEF effects since it is quite decisive on the pattern observed in the rest of the parameters. The severe performance degradation depicted in Fig. 7a points out that NOMA cannot always deliver a better performance than OMA, thus, SIC receivers should have a desirable efficiency (i.e., $1 - \epsilon$) in order to reduce the negative effects of the residual interference on the maximum throughput objective. We must also note for Fig. 7a that a higher cluster size is not beneficial after a certain value of ϵ since putting more users on the same radio resource causes higher interference due to the increasing residual interference. As $\alpha \to 1$ in Fig. 7a-7d, we observe the following behaviors: The performance gain between proposed and basic NOMA and that between basic NOMA and OMA increases monotonically. This can clearly be seen from perfect basic NOMA (—○—) and OMA (—△—) cases which are around 0.95/0.75/0.5/0.2 and 0.9/0.65/0.3/0 for α at 0/0.25/0.5/1, respectively. This is indeed because of the combination of inherited NOMA fairness and proportional fairness enforced as $\alpha \to 1$.

Moreover, increasing the performance difference between proposed and basic NOMA curves points out that higher cluster sizes more favorable as α reaches to the proportional fairness. Another important pattern to observe is that the undesirable impacts of residual interference diminish since

proposed ($-\ominus-$) and basic($-\ominus-$) NOMA gets closer to corresponding perfect cases as $\alpha \to 1$. Because the UE that contributes the total cluster sumrate is protected no more against the negative impact of the FEF as $\alpha \to 1$ optimal scheme seeks for proportional fairness not only among the clusters but also among the members of a cluster.

Fig. 8 clearly demonstrates the full benefit of allowing larger NOMA clusters. It is quite interesting that agnostic case of larger cluster sizes turns in a better performance than the perfect basic NOMA scheme of maximum throughput case in Fig. 8a. It is also clear that as α approaches the proportional fairness, the network enjoys a larger cluster size more than the maximum throughput. For instance, the ratio between the proposed and the basic NOMA is 1.5 and 1.9 for $\bar{K} = 3$ and $\bar{K} = 10$ under the maximum throughput case, respectively. On the other hand, the ratio between the proposed and the basic NOMA is 3 and 5 for $\bar{K} = 3$ and $\bar{K} = 10$ under the proportional fair objective, respectively.

The impact of UE association scheme on the network performance is demonstrated in Fig. 9. As shown in Fig. 9a, network throughput hits a peak when users are associated as per DUDe $\left(\beta = \frac{P_s}{P_m}\right)$, that monotonically degrades as $\beta \to 0$ and $\beta \to 1$ in the DUCo scheme. In particular, $\beta = 1$ loads the MBS down with the entire traffic, thus, deliver the worst performance mainly because of the deteriorated cell-edge performance and its inevitable consequence of uncancelled or residual interference to other users. Except for the proposed case, this trend also applies for other α cases. Another important pattern to observe is that negative influence of β in the network performance diminishes as $\alpha \to 1$.

Fig. 10 presents the performance trend for increasing number of SBS under different α cases. Increasing B helps the maximum throughput case due to more desirable channel gains since DUDe has a better opportunity to associate UEs with nearby BSs. However, increasing B does not show the same trend as $\alpha \to 1$ because a larger cluster size is more preferable for proportional fairness. Similarly, Fig. 11 exhibits the increasing behavior of the performance as the total number of UEs increases. Apparently, increasing the total number of UEs provide less performance increase as $\alpha \to 1$.

VIII. Conclusion

In this paper, an α-fair resource allocation and cluster formation problem is studied for DUDe HetNets under the imperfection of NOMA scheme due to the residual interference and SIC constraints. Unlike the traditional basic NOMA cluster of size two, the largest feasible cluster size is derived in the closed-form as a function cluster bandwidth, SINR requirements, and the FEF levels. Numerical results have clearly shown that a larger cluster size provides a better performance thanks to improved spectral efficiency. Furthermore, we develop a distributed cluster formation and power-bandwidth allocation framework which iteratively updates clusters, power allocations, and bandwidths. For a given bandwidth and cluster formation, optimal power control policy is derived in closed form. By extensive simulation results, we have demonstrated that delivered network performance has different trends under various network parameters.

Appendix A
Proofs for Lemma 1 and Lemma 2

Proof of Lemma 1. This proof follows from the discussion within the paragraph before Lemma 1. Exploiting the eigenvalue equation, $\mathbf{F}\boldsymbol{\nu} = \lambda_F \boldsymbol{\nu}$, we have the following set of equations

$$\frac{\lambda_F}{\bar{\Gamma}_1(\theta)}\nu_1 = \sum_{i=1}^{K} \nu_i \tag{28}$$

$$\frac{\lambda_F}{\bar{\Gamma}_i(\theta)}\nu_i = \epsilon \sum_{j=1}^{i-1} \nu_i + \sum_{k=i+1}^{K} \nu_i, i \geqslant 2, \tag{29}$$

where ν_i is the i^{th} element of the eigenvector of $\boldsymbol{\nu}$, which can be obtained recursively as follows

$$\nu_2 = \nu_1 \frac{\epsilon + \frac{\lambda_F}{\bar{\Gamma}_1(\theta)}}{1 + \frac{\lambda_F}{\bar{\Gamma}_2(\theta)}}, \; \nu_3 = \nu_2 \frac{\epsilon + \frac{\lambda_F}{\bar{\Gamma}_2(\theta)}}{1 + \frac{\lambda_F}{\bar{\Gamma}_3(\theta)}} = \tag{30}$$

$$\nu_1 \frac{\left(\epsilon + \frac{\lambda_F}{\bar{\Gamma}_1(\theta)}\right)\left(\epsilon + \frac{\lambda_F}{\bar{\Gamma}_2(\theta)}\right)}{\left(1 + \frac{\lambda_F}{\bar{\Gamma}_2(\theta)}\right)\left(1 + \frac{\lambda_F}{\bar{\Gamma}_3(\theta)}\right)}, \ldots, \nu_k = \nu_1 \prod_{i=2}^{k} \frac{\left(\epsilon + \frac{\lambda_F}{\bar{\Gamma}_{i-1}(\theta)}\right)}{\left(1 + \frac{\lambda_F}{\bar{\Gamma}_i(\theta)}\right)}.$$

Assuming a non-ideal SIC receiver, $\epsilon > 0$, $\mathbf{H}$ becomes an irreducible positive matrix. Ensuring $\lambda_F < 1$ in (30), the *Perron-Frobenius theorem* yields

$$\sum_{i=2}^{K} \prod_{j=2}^{i} \frac{(\epsilon + \frac{\lambda_F}{\bar{\Gamma}_{j-1}(\theta)})}{(1 + \frac{\lambda_F}{\bar{\Gamma}_j(\theta)})} = \frac{\lambda_F}{\bar{\Gamma}_1(\theta)} \tag{31}$$

Accounting for the feasibility condition $\lambda_F < 1$, the largest feasible cluster size falls within the range of $K_{\min} \leqslant K \leqslant K_{\max}$ where the bounds can be obtained from (31) as

$$K_{\min} = \left\lfloor \frac{\ln(\epsilon)}{\ln\left(\frac{1+\epsilon \max_i(\bar{\Gamma}_i(\theta))}{1+\max_i(\bar{\Gamma}_i(\theta))}\right)} \right\rfloor, \tag{32}$$

$$K_{\max} = \left\lfloor \frac{\ln(\epsilon)}{\ln\left(\frac{1+\epsilon \min_i(\bar{\Gamma}_i(\theta))}{1+\min_i(\bar{\Gamma}_i(\theta))}\right)} \right\rfloor. \tag{33}$$

This range tightens as the composite SINR requirements tightens and finally reduces to an exact cluster size of

$$K(\epsilon, \theta) = \left\lfloor \frac{\ln(\epsilon)}{\ln\left(\frac{1+\epsilon\bar{\Gamma}(\theta)}{1+\bar{\Gamma}(\theta)}\right)} \right\rfloor, \text{ if } \bar{\Gamma}_i(\theta) = \bar{\Gamma}(\theta), \forall i \tag{34}$$

Reverse engineering of (34) yields the attainable feasible SINR for a given cluster size as

$$\bar{\Gamma}^*(K(\epsilon, \theta)) = \frac{e^{\ln(\epsilon)/K} - 1}{\epsilon - e^{\ln(\epsilon)/K}}, \text{ if } \bar{\Gamma}_i(\theta) = \bar{\Gamma}(\theta), \forall i. \tag{35}$$

□

APPENDIX B
PROOF FOR LEMMA 3

Proof. Building upon Appendix A, optimal power levels can be derived directly by solving (18) as follows

$$p_1 = \frac{\bar{\Gamma}_1(\theta)\sigma^2}{1-\sum_{i=1}^{K}\bar{\Gamma}_i(\theta)\prod_{j=2}^{i}\left(\frac{1+\epsilon\bar{\Gamma}_{j-1}(\theta)}{1+\bar{\Gamma}_j(\theta)}\right)} \tag{36}$$

$$p_k = p_1\frac{\bar{\Gamma}_2(\theta)}{\bar{\Gamma}_1(\theta)}\prod_{j=2}^{k}\left(\frac{1+\epsilon\bar{\Gamma}_{j-1}(\theta)}{1+\bar{\Gamma}_j(\theta)}\right),\ k\geqslant 2. \tag{37}$$

which can be simplified for $\bar{\Gamma}_i(\theta) = \bar{\Gamma}(\theta), \forall i$, or $\max_i\left(\bar{\Gamma}_i(\theta)\right) = \bar{\Gamma}(\theta)$ as

$$p_k = \left(\frac{1+\epsilon\bar{\Gamma}(\theta)}{1+\bar{\Gamma}(\theta)}\right)^{k-1}\times \frac{\bar{\Gamma}(\theta)\sigma^2}{1-\left(\frac{1+\epsilon\bar{\Gamma}(\theta)}{1-\epsilon}\right)+\left(\frac{1+\epsilon\bar{\Gamma}(\theta)}{1-\epsilon}\right)\left(\frac{1+\epsilon\bar{\Gamma}(\theta)}{1+\bar{\Gamma}(\theta)}\right)^{K-1}},\ k\geqslant 2, \tag{38}$$

which follows from the fact that the second term of denominator of (36) becomes a geometric series sum by setting $\bar{\Gamma}_i(\theta) = \bar{\Gamma}(\theta), \forall i$. For $\bar{\Gamma}_i(\theta) = \bar{\Gamma}(\theta), \forall i$ or $\max_i\left(\bar{\Gamma}_i(\theta)\right) = \bar{\Gamma}(\theta)$, the largest feasible cluster size range can be obtained from (38) as $K_{\min}\leqslant K\leqslant K_{\max}$ where

$$K_{\min} = \left\lfloor 1+\frac{\ln\left(\frac{\epsilon(1+\bar{\Gamma}(\theta))}{\epsilon-1}\right)-\ln\left(\frac{\bar{\Gamma}(\theta)\sigma^2}{\bar{P}_u g_K}-\frac{1+\epsilon\bar{\Gamma}(\theta)}{1-\epsilon}\right)}{\ln\left(\frac{1+\epsilon\bar{\Gamma}(\theta)}{1+\bar{\Gamma}(\theta)}\right)}\right\rfloor, \tag{39}$$

$$K_{\max} = \left\lceil 1+\frac{\ln\left(\frac{\bar{\Gamma}(\theta)\sigma^2}{\bar{P}_u g_1}+\frac{\epsilon(1+\bar{\Gamma}(\theta))}{1-\epsilon}\right)-\ln\left(\frac{1+\epsilon\bar{\Gamma}(\theta)}{1-\epsilon}\right)}{\ln\left(\frac{1+\epsilon\bar{\Gamma}(\theta)}{1+\bar{\Gamma}(\theta)}\right)}\right\rceil. \tag{40}$$

Equations (39) and (40) are obtained by substituting (38) into $\bar{P}_u g_1\geqslant p_1$ and $\bar{P}_u g_K\geqslant p_K$, respectively, rewriting for K, and taking the natural logarithm from both sides. □

APPENDIX C
PROOF OF LEMMA 4

Proof. This appendix explains how Table I created and Lemma 4 is obtained. First, let us consider the index set of UEs who are active at maximum transmission power constraint, i.e., $\forall i\in\mathcal{I}_\lambda$. Obviously, such UEs set their power weights to unity, i.e., $\omega_i = 1, \forall i\in\mathcal{I}_\lambda$, as in the first case of (26). The optimal power weights of remaining UEs who are active at CSC, i.e., $\forall i\in\mathcal{I}_\mu$, can be directly obtained from CSCs as follows:

$$w_i\frac{h_i}{\bar{\Gamma}_i} = \epsilon\sum_{\substack{1\leqslant j<\max_i\\ j\in\mathcal{I}_\lambda}}h_j+\sum_{\substack{\max_i<j\leqslant K\\ j\in\mathcal{I}_\lambda}}h_j + \epsilon\sum_{\substack{1\leqslant j<\max_i\\ j\in\mathcal{I}_\mu}}w_j h_j+\sum_{\substack{\max_i<j\leqslant K\\ j\in\mathcal{I}_\mu}}w_j h_j+\varrho, \forall i\in\mathcal{I}_\mu \tag{41}$$

where $\max_i = \text{argmax}\{m|m\in\mathcal{I}_\mu, m<i\}$. Since UE_i, $\forall i\in\mathcal{I}_\mu$, are active at CSCs, (41) is obtained by substituting $\omega_j = 1, \forall j\in\mathcal{I}_\lambda$, and rewriting $R_i = \bar{\Gamma}_i$ for $\omega_i, \forall i\in\mathcal{I}_\mu$. Exploiting (41), $\omega_i-\omega_{\max_i}$ can be written as

$$w_i\frac{h_i}{\bar{\Gamma}_i}-w_{\max_i}\frac{h_{\max_i}}{\bar{\Gamma}_{\max_i}} = \epsilon w_{\max_i}h_{\max_i}-w_i h_i + (\epsilon-1)\sum_{\substack{\max_i<j<i\\ j\in\mathcal{I}_\lambda}}h_j,, \forall i\in\mathcal{I}_\mu \tag{42}$$

After some algebraic manipulations on 42, the first-order non-homogeneous recurrence relations with variable coefficients can be obtained as

$$w_i = w_{\max_i}\frac{h_{\max_i}\left(\epsilon+\frac{1}{\bar{\Gamma}_{\max_i}}\right)}{h_i\left(1+\frac{1}{\bar{\Gamma}_i}\right)}+\frac{(\epsilon-1)\sum_{\substack{\max_i<j<i\\ j\in\mathcal{I}_\lambda}}h_j}{h_i\left(1+\frac{1}{\bar{\Gamma}_i}\right)}, \tag{43}$$

$\forall i\in\mathcal{I}_\mu$, which is apparently in the form of $w_i = w_{\max_i}a_i+b_i$ where $a_i = \frac{h_{\max_i}\left(\epsilon+\frac{1}{\bar{\Gamma}_{\max_i}}\right)}{h_i\left(1+\frac{1}{\bar{\Gamma}_i}\right)}$ and $b_i = \frac{(\epsilon-1)\sum_{\max_i<j<i}h_j}{h_i\left(1+\frac{1}{\bar{\Gamma}_i}\right)}$. Accordingly, the recurrent relation in (43) can be rewritten as in (26) solution of which can be obtained as in ω_{mind} by following the standard procedure given in [37, Theorem 4.2]. □

REFERENCES

[1] A. Celik, R. M. Radaydeh, F. S. Al-Qahtani, A. H. A. El-Malek, and M. S. Alouini, "Resource allocation and cluster formation for imperfect NOMA in DL/UL decoupled HetNets," in *proc. IEEE Global Commun. Conf. (GLOBECOM) Workshops (GC Wkshps)*, Dec. 2017, pp. 1–6.

[2] L. Dai, B. Wang, Y. Yuan, S. Han, C. l. 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, Sept. 2015.

[3] J. G. Andrews and T. H. Meng, "Optimum power control for successive interference cancellation with imperfect channel estimation," *IEEE Trans. Wireless Commun.*, vol. 2, no. 2, pp. 375–383, Mar. 2003.

[4] N. I. Miridakis and D. D. Vergados, "A survey on the successive interference cancellation performance for single-antenna and multiple-antenna OFDM systems," *IEEE Commun. Surveys & Tutorials*, vol. 15, no. 1, pp. 312–335, First 2013.

[5] 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.

[6] Y. Liu, M. Elkashlan, Z. Ding, and G. K. Karagiannidis, "Fairness of user clustering in MIMO non-orthogonal multiple access systems," *IEEE Commun. Lett.*, vol. 20, no. 7, pp. 1465–1468, Jul. 2016.

[7] Z. Liu, L. Lei, N. Zhang, G. Kang, and S. Chatzinotas, "Joint beamforming and power optimization with iterative user clustering for MISO-NOMA systems," *IEEE Access*, vol. 5, pp. 6872–6884, 2017.

[8] M. A. Sedaghat and R. R. Müller, "On user pairing in uplink NOMA," *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 3474–3486, May 2018.

[9] J. Kang and I. Kim, "Optimal user grouping for downlink NOMA," *IEEE Wireless Commun. Lett.*, vol. 7, no. 5, pp. 724–727, Oct. 2018.

[10] M. S. Elbamby, M. Bennis, W. Saad, M. Debbah, and M. Latva-aho, "Resource optimization and power allocation in in-band full duplex-enabled non-orthogonal multiple access networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2860–2873, Dec. 2017.

[11] S. Ali, E. Hossain, and D. I. Kim, "Non-orthogonal multiple access (NOMA) for downlink multiuser MIMO systems: User clustering, beamforming, and power allocation," *IEEE Access*, vol. 5, pp. 565–577, 2017.

[12] 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, 2016.

[13] J. Ding, J. Cai, and C. Yi, "An improved coalition game approach for mimo-noma clustering integrating beamforming and power allocation," *IEEE Trans.Vehicular Technol.*, vol. 68, no. 2, pp. 1672–1687, Feb. 2019.

[14] J. Cui, Z. Ding, P. Fan, and N. Al-Dhahir, "Unsupervised machine learning-based user clustering in millimeter-wave-noma systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 11, pp. 7425–7440, No.v 2018.
[15] Y. Tsai and H. Wei, "Quality-balanced user clustering schemes for non-orthogonal multiple access systems," *IEEE Commun. Lett.*, vol. 22, no. 1, pp. 113–116, Jan. 2018.
[16] 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.
[17] Z. Wei, D. W. K. Ng, J. Yuan, and H. Wang, "Optimal resource allocation for power-efficient MC-NOMA with imperfect channel state information," *IEEE Trans. Commun.*, vol. 65, no. 9, pp. 3944–3961, Sep. 2017.
[18] L. Xiao, Y. Li, C. Dai, H. Dai, and H. V. Poor, "Reinforcement learning-based NOMA power allocation in the presence of smart jamming," *IEEE Trans. Vehicular Technol.*, vol. 67, no. 4, pp. 3377–3389, Apr. 2018.
[19] H. Tabassum, E. Hossain, and J. Hossain, "Modeling and analysis of uplink non-orthogonal multiple access in large-scale cellular networks using poisson cluster processes," *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3555–3570, Aug. 2017.
[20] 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.
[21] N. Zhang, J. Wang, G. Kang, and Y. Liu, "Uplink nonorthogonal multiple access in 5G systems," *IEEE Commun. Lett.*, vol. 20, no. 3, pp. 458–461, Mar. 2016.
[22] J. Choi, "On power and rate allocation for coded uplink NOMA in a multicarrier system," *IEEE Trans. Commun.*, vol. 66, no. 6, pp. 2762–2772, Jun. 2018.
[23] K. N. Pappi, P. D. Diamantoulakis, and G. K. Karagiannidis, "Distributed uplink-NOMA for cloud radio access networks," *IEEE Commun. Lett.*, vol. 21, no. 10, pp. 2274–2277, Oct. 2017.
[24] 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.
[25] A. Celik, F. S. Al-Qahtani, R. M. Radaydeh, and M. S. Alouini, "Cluster formation and joint power-bandwidth allocation for imperfect NOMA in DL-HetNets," in *proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2017, pp. 1–6.
[26] A. Celik, M. Tsai, R. M. Radaydeh, F. S. Al-Qahtani, and M. Alouini, "Distributed cluster formation and power-bandwidth allocation for imperfect NOMA in DL-HetNets," *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1677–1692, Feb. 2019.
[27] F. Boccardi, J. Andrews, H. Elshaer, M. Dohler, S. Parkvall, P. Popovski, and S. Singh, "Why to decouple the uplink and downlink in cellular networks and how to do it," *IEEE Commun. Mag.*, vol. 54, no. 3, pp. 110–117, Mar. 2016.
[28] A. Celik, R. M. Radaydeh, F. S. Al-Qahtani, and M.-S. Alouini, "Joint interference management and resource allocation for device-to-device (D2D) communications underlying downlink/uplink decoupled (DUDe) heterogeneous networks," in *proc. IEEE Intl. Conf. Commun.(ICC)*, May 2017.
[29] A. Hasan and J. Andrews, "Cancellation error statistics in a power-controlled cdma system using successive interference cancellation," in *8th IEEE International Symposium on Spread Spectrum Techniques and Applications*, Aug. 2004, pp. 419–423.
[30] J. G. Andrews, "Interference cancellation for cellular systems: a contemporary overview," *IEEE Wireless Commun.*, vol. 12, no. 2, pp. 19–29, Apr. 2005.
[31] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Trans. Netw.*, vol. 8, no. 5, pp. 556–567, Oct. 2000.
[32] E. Altman, K. Avrachenkov, and A. Garnaev, "Generalized α-fair resource allocation in wireless networks," in *2008 47th IEEE Conference on Decision and Control*, Dec. 2008, pp. 2414–2419.
[33] A. Agrawal, J. G. Andrews, J. M. Cioffi, and T. Meng, "Iterative power control for imperfect successive interference cancellation," *IEEE Trans. Wireless Commun.*, vol. 4, no. 3, pp. 878–884, May 2005.
[34] W. Kahan, "Ieee standard 754 for binary floating-point arithmetic," *Lecture Notes on the Status of IEEE*, vol. 754, no. 94720-1776, p. 11, 1996.
[35] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge university press, 1990.
[36] E. K. Chong and S. H. Zak, *An introduction to optimization*. John Wiley & Sons, 2013, vol. 76.
[37] A. Jensen, "Lecture notes on difference equations," Jul. 2011.

Abdulkadir Çelik (S'14-M'16) received the B.S. degree in electrical-electronics engineering from Selçuk University, Konya, Turkey in 2009, the M.S. degree in electrical engineering in 2013, the M.S. degree in computer engineering in 2015, and the Ph.D. degree in co-majors of electrical engineering and computer engineering in 2016, all from Iowa State University, Ames, IA, USA. He is currently a postdoctoral research fellow at Communication Theory Laboratory of King Abdullah University of Science and Technology (KAUST). His current research interests include but not limited to 5G and beyond, wireless data centers, UAV assisted cellular and IoT networks, and underwater optical wireless communications, networking, and localization.

Ming-Cheng Tsai (S'18) was born in Fujian, China. He received his B.E. degree in electrical engineering from the National Taipei University of Technology, Taipei, Taiwan, in 2015. From 2015 to 2018, he was a student of Communication Engineering at the National Tsinghua University. He is currently pursuing his M.S./Ph.D. degree in Electrical Engineering at King Abdullah University of Science and Technology. His research interests include device to device communication, digital communication, and error correcting codes.

Redha M. Radaydeh (S'05-M'07-SM'13) was born in Irbid, Jordan, on November 12, 1978. He received the B.S. and M.S. degrees from Jordan University of Science and Technology (JUST), Irbid, in 2001 and 2003, respectively, and the Ph.D. degree from University of Mississippi, Oxford, MS, USA, in 2006, all in electrical engineering. He worked at JUST, King Abdullah University of Science and Technology (KAUST), Texas A& M University at Qatar (TAMUQ), and Alfaisal University as an associate professor of electrical engineering. He also worked as a remote research scientist with KAUST and was a visiting researcher with Texas A& M University (TAMU), College Station, TX. Currently, he is a faculty member with the electrical engineering program at Texas A& M University-Commerce (TAMUC), Commerce, TX. His research interests include broad topics on wireless communications, and design and performance analysis of wireless networks.

Fawaz S. Al-Qahtani (M'10) received the B.Sc. in electrical engineering from King Fahad University of Petroleum and Minerals (KFUPM), Saudi Arabia in 2000 and M.Sc. in Digital Communication Systems from Monash University, Melbourne, Australia in 2005, and Ph.D. degree in Electrical and Computer Engineering, from RMIT University, Australia in December, 2009. He is currently working as IP commercialization manager for ICT portfolio at Research, Development, and Innovation (RDI) in Qatar Foundation, Doha, Qatar. From May 2010 to August 2017, he was research scientist with Texas A& M University at Qatar, Education City, Doha, Qatar. His research has been sponsored by Qatar National Research Fund (QNRF). He was awarded of JSERP and NPRP projects. His current research interests include channel modeling, applied signal processing, MIMO communication systems, cooperative communications, cognitive radio systems, free space optical, physical layer security, and device-to-device communication. He is the author or co-author of 100 technical papers published in scientific journals and presented at international conferences.

Mohamed-Slim Alouini (S'94-M'98-SM'03-F'09) was born in Tunis, Tunisia. He received the Ph.D. degree in Electrical Engineering from the California Institute of Technology (Caltech), Pasadena, CA, USA, in 1998. He served as a faculty member in the University of Minnesota, Minneapolis, MN, USA, then in the Texas A&M University at Qatar, Education City, Doha, Qatar before joining King Abdullah University of Science and Technology (KAUST), Thuwal, Makkah Province, Saudi Arabia as a Professor of Electrical Engineering in 2009. His current research interests include the modeling, design, and performance analysis of wireless communication systems.