Cluster Formation and Joint Power-Bandwidth Allocation for Imperfect NOMA in DL-HetNets

Abdulkadir Celik, Fawaz S. Al-Qahtani, Redha M. Radaydeh, and Mohamed-Slim Alouini

Abstract—Non-orthogonal multiple access (NOMA) has recently drawn attentions on its ability to fairly serve multiple users on the same radio resource with a desirable performance. However, achievable NOMA gain is primarily limited by channel gain disparity and successive interference cancellation (SIC) receiver characteristics. Accordingly, we introduce an imperfect SIC receiver model considering the power disparity and sensitivity constraints, delay tolerance, and residual interference due to detection and estimation errors. Then, a generic cluster formation (CF) and Power-Bandwidth Allocation (PBA) is formulated as a mixed-integer non-linear programming (MINLP) problem for downlink (DL) heterogeneous networks (HetNets). After dividing the MINLP problem into mixed-integer and non-linear subproblems, we first transform CF into a multi-partite matching, which is solved sequentially using bi-partite matching techniques. For sumrate maximization, max-min fairness, and energy & spectrum efficiency objectives, we secondly put highly non-convex joint PBA into a convex form using geometric programming (GP). Extensive simulations unleash the potential of NOMA to handle large number of users, traffic offloading, and user fairness.

I. INTRODUCTION

ULTRA-DENSE networks have been considered as a promising solution since network densification has the ability to boost network coverage and capacity, while reducing operational and capital expenditures of fifth generation (5G) networks [1]. However, traditional HetNets dedicate radio resources to a certain user either in time or frequency domains, i.e., orthogonal multiple access (OMA) where the number of connected/served users at a given time instant is strictly limited by the resource availability. Also considering the expected explosive number of devices, required massive connectivity necessitates more spectrum efficient access schemes with extended coverage.

In this regard, NOMA has recently attracted attentions by allowing multiple users to share the same time and frequency resources. In particular, power domain NOMA (PD-NOMA) can serve multiple users at different power levels by ensuring that some users can cancel some of the interference out before decoding their own signal. In this way, high channel gain users with low power transmission can improve their performance by canceling low channel gain users' interference. As low channel gain users cannot cancel high channel gain users, their performance loss is mitigated by high power levels. Therefore, performance gain achieved by NOMA is primarily determined by reception power and channel gain disparity [2]. Therefore, one of the foremost challenges of the NOMA

comes from combinatoric and non-linear nature of the CF and PBA problems, which is the main focus of this paper.

Recent efforts on CF and power allocation for PD-NOMA can be exemplified as follows: PD-NOMA was adopted by 3GPP long-term evolution advanced (3GPP LTE-A) networks for multiuser superposition transmission (MUST) [3]. The results showed that the MUST can increase the system capacity and improve the user experience significantly. The work in [2] developed new analytical results for ergodic sum rate and outage probability of a cellular DL scenario with randomly deployed users. It was shown that NOMA can achieve substantial performance enhancement in terms of ergodic sum rates, however, the outage performance depends on the users' targeted data rates and allocated powers. User fairness in DL-NOMA systems has been investigated in [4] considering perfect channel state information channel state information (CSI), and average CSI feedback. The impact of user pairing was investigated in [5] for a two-user DL-NOMA system with fixed power allocation and cognitive radio inspired power allocation. User clustering and resource allocation of DL/UL decoupled HetNets is studied in [6] where clusters are formed using blossom algorithm. Moreover, it has been shown that while using DL-NOMA decoding error does not provide any performance enhancement over OMA, using an inverse decoding order can achieve NOMA gain which is upper bounded by channel gain disparity. Joint power and channel allocation for NOMA system has been investigated in [7] wherein near optimal solution was proposed by combining Lagrangian duality and dynamic programming. The work in [8] incorporated multiple antennas with NOMA systems, and proposed memorization-maximization method for optimizing sum rates of DL-NOMA systems. Authors of [9] characterized the ergodic sum capacity maximization problem of a two-user multiple input multiple output (MIMO)-NOMA system under statistical CSI with the total power constraint and minimum rate constraint for the weak user. Recently, [10] first groups users into a single cluster, and then optimizes their respective power allocation. Aside from works above, our main contributions in this paper can be enlisted as follows:

- 1) In practice, constraints and imperfections of SIC receivers constitute a limiting factor on achievable gain by NOMA. For a successful interference cancellation (IC), SIC receivers must observe a certain signal power disparity which is mainly determined by the hardware sensitivity. Furthermore, decoding process imperfections leaves some residual interference at the very end. Thus, we first model a practical NOMA receiver and show impacts of constraints and imperfections on achievable NOMA performance.

- 2) After formulating the optimal CF and PBA as an MINLP problem, we develop a solution methodology by considering

A. Celik, R. M. Radaydeh, and M-S. Alouini are with Computer, Electrical, and Mathematical Sciences and Engineering Division at King Abdullah University of Science and Technology (KAUST), Thuwal, 23955-6900, KSA. F. S. Al-Qahtani is with Department of Electrical and Computer Engineering, Texas A&M University, Doha, Qatar.

mixed-integer CF and non-linear PBA parts separately. A generic CF scheme is obtained by creating a *multi-partite graph* (MPG) via partitioning users with respect to their channel gains. Weighting metrics of the MPG is determined to consider both channel gain disparity and QoS demands of users. Resulted *weighted multi-partite matching* (WMM) is then solved sequentially using *weighted bi-partite matching* (WBM) methods.

3) For given CFs, we finally show how PBA can be put in standard GP form for energy and spectrum efficiency, network sumrate maximization, and max-min fairness objectives. Then, we investigate impacts of traffic offloading, user density, and different cluster sizes and compare NOMA and OMA schemes under aforementioned objectives.

The remainder of the paper is organized as follows: Section II introduces the system model along with constraints and imperfections of SIC. Then, Section III first formulates the optimal CF and PBA problem and provide developed solution methodology. Numerical results are presented in Section IV and Section V concludes the paper with a few remarks.

II. SYSTEM MODEL

We consider DL transmission of a 2-tiered HetNet where each tier represents a particular cell class, i.e., tier-1 consists of a single macrocell and tier-2 comprises of smallcells. The spatial distribution of SBSs and user equipments (UEs) follow Poisson point process (PPP) Φ_s of density λ_s and Φ_u of density λ_u , respectively. Denoting the number of SBSs as S , the index set of all cells/BSs is denoted by $\mathcal{C} = \{c | 0 \leq c \leq S\}$ where $c = 0$ and $0 < c \leq S$ represent the macrocell and smallcells, respectively. We note that the terms BS, cell and their indices will be used interchangeably throughout the paper. Maximum transmission powers of BSs are generically denoted as P_c which equals to P_m and P_s for MBS and SBSs, respectively.

For a given user-cell association, $\mathcal{U}_c$ denotes the index set of U_c UEs associated with BS_c . UEs are assumed to be equipped with a single antenna and single SIC receiver. Furthermore, index set of all $U = \sum_c U_c$ UEs is given as $\mathcal{U} = \bigcup_c \mathcal{U}_c$. $\mathcal{U}_c$ is partitioned into clusters such that set of users within cluster r is symbolized as $\mathcal{K}_c^r \triangleq \{i | \alpha_{c,r}^i \in \{0, 1\}\}$ where $\alpha_{c,r}^i$ is a binary indicator for the association of UE_i with $\mathcal{K}_c^r$. Denoting the set of all clusters as $\mathcal{R}$, cluster r of cell c is allowed to utilize $0 \leq \theta_c^r \leq 1$, $\sum_{c,r} \theta_c^r \leq 1$, portion of the entire DL bandwidth, B , as per NOMA principles.

In DL-NOMA, while BSs broadcast the superposition of messages for its subscribed members. Therefore, received signal by $\text{UE}_i \in \mathcal{K}_c^r$ can be expressed as

$$z_{c,r}^i = g_c^i (A_{c,r}^i \alpha_{c,r}^i y_i + \underbrace{\sum_{j \neq i} A_{c,r}^j \alpha_{c,r}^j y_j}_{\text{Intra-cluster Intf.}}) + n_c^r \quad (1)$$

where $A_{c,r}^i = (P_c \omega_{c,r}^i)^{1/2}$ is the signal amplitudes for UE_i , $\omega_{c,r}^i$ is the transmission power weight of BS_c for $\text{UE}_i \in \mathcal{K}_c^r$ such that $\sum_{i \in \mathcal{K}_c^r} \omega_{c,r}^i \leq 1$, $\forall c$, y_i is the signal intended for UE_i , and $n_c^r \sim \mathcal{N}(0, \sigma_{c,r}^2)$ is the additive white Gaussian noise with variance $\sigma_{c,r}^2 = N_0 \theta_c^r B$ where N_0 is the thermal noise power

Fig. 1: Illustration of user clustering and traffic offloading.

spectral density. The composite channel gain, g_j^i , between a generic transmitter node j and a generic receiver node i is given as

$$g_j^i = K_j^i \delta_{j,i}^{-\eta_j^i} 10^{\xi_j^i/10} |h_j^i|^2 \quad (2)$$

where K_j^i is a constant related to antenna parameters, $\delta_{j,i}$ is the distance between the nodes, η_j^i is the path loss exponent, $10^{\xi_j^i/10} \sim \ln \mathcal{N}(0, \sigma_j^i)$ represents the log-normally distributed shadowing, ξ_j^i is a normal random variable representing the variation in received power with a standard deviation of σ_j^i , i.e., $\xi_j^i \sim \mathcal{N}(0, \sigma_j^i)$, and h_j^i is the complex channel fading coefficient.

Among other interference cancellation (IC) methods, successive IC (SIC) is regarded as the most effective IC technique in terms of robustness and applicability in practice [11]. In the DL-NOMA, BSs broadcast the superposed signals with low power level for high channel gain users and high power level signals for low channel gain users. To extract desired signal from composite signal, the SIC receiver first decodes the strongest interference, then re-generates the transmitted signal by re-encoding and re-modulation, and finally subtracts it from the received composite signal, which is repeated for succeeding interference components. Let us now focus on a single cluster of BS_c $\mathcal{K}_c^r = \{i | g_c^{i-1} \geq g_c^i \geq g_c^{i+1}, \forall i\}$ which is allocated to subcarrier r with bandwidth θ_c^r . Fig. 1 illustrates an exemplary MBS cluster, $\mathcal{K}_0^r = \{1, 2, \dots, 10\}$. As per SIC principles, NOMA allocates transmission power weights as $\omega_c^{i-1} < \omega_c^i < \omega_c^{i+1}$, $\forall i \in \mathcal{K}_c^r$, and decoding order set for UE_i is obtained as

$$\underbrace{p_{K_c^r}^i > \dots > p_{i+1}^i}_{\text{Higher Rank Decoding Order: } \mathcal{O}_i^h} > p_i^i > \underbrace{p_{i-1}^i > \dots > p_1^i}_{\text{Lower Rank Decoding Order: } \mathcal{O}_i^l} \quad (3)$$

where K_c^r is the cluster size, p_j^i is the power of $\text{UE}_j \in \mathcal{K}_c^r$'s message received by UE_i , $\mathcal{O}_i^h = \{i+1, \dots, K_c^r\}$ is the higher rank decoding order set, and $\mathcal{O}_i^l = \{1, \dots, i\}$ is the lower rank decoding order set for UE_i . UE_i can only cancel the interference induced from higher rank members, while interference from lower rank members cannot be decoded as they are weaker than the desired signal. Furthermore, the

hardware sensitivity of the SIC receiver requires a minimum signal-to-noise-ratio (SNR) to distinguish signal from noise. Therefore, *power disparity and sensitivity constraints* (PDSCs) can be expressed in linear scale as [10]

$$p_j^i - \sum_{k=1}^{j-1} p_k^i \geq p_\Delta, \quad \forall \text{ UE}_i, \forall \text{ UE}_j \in \mathcal{O}_i^h \quad (4)$$

where p_Δ denotes the hardware sensitivity. The intuition behind (4) is that during the IC process of $\text{UE}_j \in \mathcal{O}_i^h$, receiver observes undecoded signals of $\text{UE}_k \in \mathcal{O}_i^h$, $k < j$, as noise. Moreover, a non-ideal SIC observes some residual interference after SIC due to detection errors and CSI imperfections. Accordingly, a generic signal-to-interference-plus-noise-ratio (SINR) representation of the imperfect SIC receiver can be given by

$$\gamma_{c,r}^i = \frac{p_{c,r}^{i,i}}{\sum_{j \in \hat{\mathcal{O}}_i^h} \epsilon_i p_{c,r}^{j,i} + \sum_{k \in \mathcal{O}_i^h \setminus \hat{\mathcal{O}}_i^h} p_{c,r}^{k,i} + \sum_{l \in \mathcal{K}_c^r} p_{c,r}^{l,i} + \sigma_{c,r}^2} \quad (5)$$

where $0 \leq \epsilon_i \leq 1$ is the residual interference factor, $p_{c,r}^{i,i} = P_c \omega_{c,r}^i g_c^i \alpha_{c,r}^i$, and $p_{c,r}^{j,i} = P_c \omega_{c,r}^j g_c^j \alpha_{c,r}^j$. Denoting the affordable number of ICs by highest channel gain user as L_c , $\hat{\mathcal{O}}_i^h$ is the set of first $\min(L_c, |\mathcal{O}_i^h|)$ most dominant interference elements of $\mathcal{O}_i^h$. The first, second, and third terms in denominator represent the residual interference after cancellation, uncanceled interference due to delay constraint, and uncanceled lower rank interference, respectively. Hence, the capacity of UE_i is given by

$$\mathbb{C}_i = \sum_{c,r} \alpha_{c,r}^i \theta_c^r B \log_2(1 + \gamma_{c,r}^i) \quad (6)$$

III. CLUSTER FORMATION AND POWER-BANDWIDTH ALLOCATION

In HetNets, $\text{UE} \leftrightarrow \text{BS}$ associations are determined by the received signal strength (RSS) information. Thus, MBSs are loaded with significant amount of traffic since their high transmission power yields large number of user association. In order to offload data traffic from MBSs to SBSs, $\text{UE} \leftrightarrow \text{BS}$ association is done by introducing a bias factor, $0 \leq \beta \leq 1$. While $\beta = 0$ is the setting where entire traffic is offloaded to SBSs (i.e., there is no serving MBS), $\beta = 1$ is the case where no traffic offload occurs. A simple case of traffic offloading is illustrated in Fig. 1 where cell-edge macrocell UEs located at the orange colored offloading zones of the SBSs are associated with the SBSs. As channel gain disparity is desirable for a better NOMA performance, SBSs can take care of the offloaded UEs by treating them as cell-edge users and pairing them with their original subscribers in their coverage area which is shown in red color. In this case, original subscribers of smallcell enjoys the interference-free (i.e., orthogonalized) performance while offloaded SUEs can still obtain a good service. This is a main result of the inherent fairness feature of NOMA as low channel gain users are compensated with relaxed interference and high power allocation, on the other side, high channel gain users with low power allocation are compensated with IC opportunities.

A. Optimal Problem Formulation

In a large-scale network, determining optimal values for integer valued cluster number/sizes and binary valued user-cluster associations induces impractical computational complexity. Therefore, we relax this combinatoric problem by making following assumptions: 1) Each cluster have a dedicated band so that inter-cluster interference is avoided within a macrocell coverage area, and 2) The cluster size is determined by orthogonalizing the highest channel gain user.

$$\begin{aligned} \mathbf{P}_o : & \max_{\alpha, \omega, \theta} \quad \mathbb{O}(\alpha, \omega, \theta) \\ C_1 : & \text{s.t.} \quad \sum_r \alpha_{c,r}^u = 1, \quad \forall c, u \\ C_2 : & \quad \sum_u \alpha_{c,r}^u \leq L_c, \quad \forall c, r \\ C_3 : & \quad \sum_{r,u} \omega_{c,r}^u \leq 1, \quad \forall c \\ C_4 : & \quad \sum_{c,r} \theta_c^r \leq 1, \\ C_5 : & \quad \frac{p_\Delta}{p_{c,r}^{j,u} - \sum_{k=1}^{j-1} p_{c,r}^{k,u}} \leq 1, \quad \forall c, r, \forall u, j \in \mathcal{O}_u^h \\ C_6 : & \quad q_u \leq \mathbb{C}_u, \quad \forall u \\ C_7 : & \quad \alpha_{c,r}^u \in \{0, 1\}, \quad 0 \leq \theta_c^r \leq 1, \quad 0 \leq \omega_{c,r}^u \leq \alpha_{c,r}^u \end{aligned}$$

For a generic objective function, $\mathbb{O}(\alpha, \omega, \theta)$, joint CF and PBA problem can be formulated as in $\mathbf{P}_o$ where vectorized form of power allocation variables is denoted as $\omega = [\omega_1^u, \dots, \omega_c^u, \dots, \omega_{S+1}^u]$, $\omega_c^u = [\omega_{c,1}^u, \dots, \omega_{c,r}^u, \dots, \omega_{c,R_c}^u]$, $\omega_{c,r}^u = [\omega_{c,r}^1, \dots, \omega_{c,r}^u, \dots, \omega_{c,r}^{U_c}]$. α follows the same notation of ω . Likewise, vectorized form of bandwidth allocation variables is given by $\theta = [\theta_1, \dots, \theta_c, \dots, \theta_{S+1}]$ where $\theta_c = [\theta_c^1, \dots, \theta_c^r, \dots, \theta_c^{R_c}]$. In $\mathbf{P}_o$, C_1 ensures that a UE is assigned to at most one cluster and C_2 limits the number of UEs within a cluster by L_c , which is a design parameter. C_3 and C_4 constraint the total power and bandwidth allocation weights to unity. PDSCs and QoS requirements are introduced in C_5 and C_6 , respectively. C_8 defines optimization variables' domains where the power allocation for UE_u on cluster r to zero ($0 \leq \omega_{c,r}^u \leq \alpha_{c,r}^u$) if $\text{UE}_u \notin \mathcal{K}_c^r$. $\mathbf{P}_o$ is apparently an MINLP problem which requires impractical time complexity even for moderate sizes of the network. As a fast yet high performance suboptimal solution methodology is of the essence to employ NOMA in practice, we develop a solution methodology by decoupling this hard problem into subproblems.

B. Cluster Formation Using Multi-partite Matching

CF is a critical step to maximize the achievable performance gain by NOMA. Hence, designing a clustering technique which accounts for channel gain disparity and QoS requirements of cluster members is utmost importance. To be more specific, let us consider the CF of UEs associated with BS_c . For a given L_c , maximum size of clusters is limited by $L_c + 1$ as the highest channel gain UE within a cluster can cancel interference of at most L_c UEs.

Accordingly, user index set of BS_c , $\mathcal{U}_c = \{i | g_c^i > g_c^{i+1}, 1 \leq i \leq U_c\}$, can be partitioned into $L_c + 1$ disjoint channel gain levels, i.e.,

$$\mathcal{P}_c^\ell = \{i | i \in \mathcal{U}_c; (\ell - 1)R_c + 1 \leq i \leq \ell R_c\}, \text{ for } \ell < L_c$$

$$\mathcal{P}_c^{L_c} = \{i | i \in \mathcal{U}_c; (L_c - 1)R_c + 1 \leq i \leq U_c\}$$

where $R_c = \lceil U_c/L_c \rceil$ is the number of clusters within $\mathcal{U}_c$ as a function of the cluster size and $\mathcal{P}_c^{\ell'} \cap \mathcal{P}_c^\ell = \emptyset, \forall \ell' \neq \ell$. That is, intra-partition and inter-partition channel gains are in descending order (i.e., $g_c^i \geq g_c^j, i, j \in \mathcal{P}_c^\ell, \forall i < j$), and the lowest channel gain within $\mathcal{P}_c^\ell$ is higher than all channel gains within $\mathcal{P}_c^{\ell'}, \forall \ell' > \ell$. One can also think of partitions as UEs falling into non-overlapping spatial ring zones as shown in Fig. 1. Based on MPG depicted in Fig. 2, CF subproblem can

Fig. 2: Multi-partite Matching Representation for CF

be formulated as a WMM. As it is a hard problem, WMM can be solved in a sequential manner such that sequence $\ell, 0 \leq \ell \leq L_c$ is a WBM between partitions $\mathcal{P}_c^\ell$ and $\mathcal{P}_c^{\ell+1}$.

Denoting the cluster set at the end of sequence ℓ as $\mathcal{K}_c^r(\ell), 1 \leq r \leq R_c, \mathcal{K}_c^1(L_c) = \{1, R_c, \dots, r, 1\}$ represents the final clustering set, for example, the sequential formation of first cluster $\mathcal{K}_c^1$ is highlighted in Fig. 2 with red colors. Considering the MPG as a matrix of size $R_c \times L_c + 1$ where columns are represented by partitions, $H_c \in \mathbb{R}^{R_c \times L_c + 1}$ and $Q_c \in \mathbb{R}^{R_c \times L_c + 1}$ represent the channel gain and QoS requirements, respectively. The edge weight from r^{th} element of $\mathcal{P}_c^\ell$ to s^{th} element of $\mathcal{P}_c^{\ell+1}$ can be given by

$$\mathcal{E}_{c,\ell}^{r,s} = \frac{h_\ell^r/h_{\ell+1}^s}{|q_\ell^r - q_{\ell+1}^s|} \quad (7)$$

which makes WBM favor for new cluster members with high channel gain disparity and close QoS requirements. Note that each matching sequence is in the form of rectangular assignment problem (RAP) which is generally solved by Munkres Algorithm in cubic order [12].

C. Joint Bandwidth and Power Allocation

For a given CF $\bar{\alpha}$, we now focus on joint PBA problem

$$\mathbf{P}_o^i(\omega, \theta | \bar{\alpha}) : \min_{\omega, \theta} \mathbb{O}_i(\omega, \theta) \text{ s.t. } C_3 - C_7$$

where we consider following objectives

1) Energy-Spectrum Cost Minimization:

$$\mathbb{O}_1(\omega, \theta) = \left(\sum_{c,r,u} \omega_{c,r}^u \right)^a \left(\sum_{c,r} \theta_c^r \right)^b \text{ where } 0 \leq$$

$a, b \leq 1$ can be manipulated to obtain energy & spectrum saving, spectrum saving, and energy saving by setting (a, b) to $(1, 1)$, $(0, 1)$, and $(1, 0)$, respectively.

2) *Network Sumrate Maximization*: $\mathbb{O}_2(\omega, \theta) = \sum_{c,r,u} \mathcal{C}_u$

3) *Max-Min Fairness*: $\mathbb{O}_3(\omega, \theta) = \min(\mathcal{C}_u)$

As being in the class of nonlinear problems, GP is a powerful and high speed optimization tool in order to handle communication network problems. GP exploits many useful theoretical and computational properties of monomials and posynomials which are defined as follows [13]

Definition 1. A function $f : \mathbb{R}_{++}^n \rightarrow \mathbb{R}$ is defined as a monomial such that

$$f(\mathbf{x} | \kappa, \mathbf{e}) = \kappa x_1^{e(1)} x_2^{e(2)} \dots x_n^{e(n)} = \kappa \prod_j x_j^{e(j)} \quad (8)$$

where the multiplicative constant $\kappa \geq 0$ and the exponential constants $e(j) \in \mathbb{R}, j = 1, 2, \dots, n$. On the other hand, positive sum of monomials is defined as a posynomial

$$g(\mathbf{f}) = \sum_i f_i(\mathbf{x}_i | \kappa_i, \mathbf{e}_i) = \sum_i \kappa_i \prod_j x_j^{e_i(j)} \quad (9)$$

where $\kappa_i \geq 0$ and the exponential constants $e_i(j) \in \mathbb{R}, j = 1, 2, \dots, n$. Composition of posynomials are further referred to as generalized posynomials which can generally be handed by GP solvers as posynomials.

Even though standard form of GP problems (GPPs) are non-convex due to the non-convexity of posynomials, it can easily be converted into an equivalent convex problem based on following features of posynomials: 1) Posynomials are closed under addition, multiplication, and positive scaling, 2) If f is a monomial and g is a posynomial, then g/f is also a posynomial, and 3) Posynomials can be converted into a convex form with logarithmic change of functions, variables, and multiplicative constants [14]. Recent solution methods can solve even large-scale GPs extremely efficiently and reliably [13]. To give a rough idea of the current state of the art, a small desktop computer can solve a typical sparse GP with 10^4 variables and 10^6 constraints in under a minute using standard interior-point algorithms [15]. In what follows, we explain how $\mathbf{P}_o^i$ s can be put into standard GP form starting from their common set of constraints $C_3 - C_7$.

Total power and bandwidth allocation constraints in C_3 and C_4 are simply sum of variables and can easily be recognized as posynomials. Received signal powers can also be shown as monomials, i.e.,

$$p_{c,r}^{i,j}(\omega | \mathbf{e}, \bar{\alpha}) = \kappa_{c,r}^j \prod_{\forall c',r',j'} (\omega_{c',r'}^{j'})^{e_{c',r'}^{j'}} \quad (10)$$

where $\mathbf{e}$ is the vector of exponential with similar notation as in $\omega, \kappa_{c,r}^{i,j} = P_c g_c^i \bar{\alpha}_{c,r}^j$ and $e_{c',r'}^{j'} = 1$ if $(c', r', j') = (c, r, j)$, otherwise, exponentials are zero valued. Please note that $\omega_{c,r}^u \leq \alpha_{c,r}^u$ in C_7 is embedded in the multiplicative constant of the monomial in (10). Hence, the nominator of the SINR expression in (5) is a monomial while the interference power in the denominator is a posynomial. Noting that $\sigma_{c,r}^2 = N_0 B \theta_c^r$ is a monomial, inverse SINR (ISINR) $\rho_{c,r}^u \triangleq 1/\gamma_{c,r}^u$ is also a

posynomial since it is ratio of a posynomial to a monomial. PDSCs in C_5 , on the other hand, can be put into a posynomial form as $\frac{p_{\Delta} + \sum_{k=1}^{j-1} p_{c,r}^{k,u}}{p_{c,r}^{j,u}} \leq 1$ which is nothing but a posynomial to monomial ratio. Assuming that cluster are formed to ensure high SINR regime, we have $\log_2(1 + \gamma_{c,r}^i) \simeq \log_2(\tilde{\gamma}_{c,r}^i)$. Therefore, QoS constraints in C_6 can be written in the form of a generalized posynomial as $\mathbb{k}_u^{1/\theta_c^r} \rho_{c,r}^u \leq 1$ where $\mathbb{k}_u = 2^{q_u/B}$.

In P_o^1 , the objective is a posynomial for integer valued (a, b) (i.e., $a \in \{0, 1\}$ and $b \in \{0, 1\}$) as it is a ratio of posynomials. For real valued factors, however, posynomials to real valued powers is not suitable for GPPs. Fortunately, this can be handled by introduction of scalar auxiliary variables ϕ and ψ such that the objective is changed to ratio of monomials $\mathbb{O}_1 = \phi^a / \psi^b$ with additional posynomial constraints $\sum_{c,r,u} \omega_{c,r}^u \leq \phi$ and $\sum_{c,r} \theta_c^r \leq \psi$. For P_o^2 , network sumrate maximization $\mathbb{O}_2(\omega, \theta)$ can be transformed into product of posynomials as follows

$$\begin{aligned} \max_{\omega, \theta} \sum_{c,r,u} \theta_c^r \log_2(1 + \gamma_{c,r}^u) &\stackrel{(*)}{\approx} \max_{\omega, \theta} \sum_{c,r,u} \log_2((\tilde{\gamma}_{c,r}^u)^{\theta_c^r}) \\ &\stackrel{(**)}{=} \min_{\omega, \theta} \prod_{c,r,u} (\rho_{c,r}^u)^{\theta_c^r} \stackrel{(***)}{\leq} \min_{\omega, \theta, \varrho} \prod_{c,r,u} (\varrho_{c,r}^u)^{\theta_c^r} \end{aligned} \quad (11)$$

where the approximation $(*)$ is based on high SINR assumption. Due to the monotonicity of the logarithm, maximization of $\log(f(\cdot))$ is equivalent to maximization of $f(\cdot)$ which yields $(**)$ by changing the objective from maximization of SINRs to minimization of ISINRs. As mentioned before, non-integer powered posynomials can be handled by introduction of auxiliary variables and constraints, i.e., $\rho_{c,r}^u \leq \varrho_{c,r}^u$ which naturally yields the transformation of objective as in $(***)$, $\min_{\omega, \theta, \varrho} \prod_{c,r,u} (\varrho_{c,r}^u)^{\theta_c^r}$. Finally, P_o^3 can be adjusted to GP framework by maximizing a scalar auxiliary variable φ which is set as a common QoS threshold for all users, i.e., $\mathbb{C}_u > \varphi$, $\forall u$. Please note that while ω and θ can jointly be optimized in $\mathbb{O}_1$, it is not possible for $\mathbb{O}_2$ and $\mathbb{O}_3$ since the base and power of $(\varrho_{c,r}^u)^{\theta_c^r}$ are both optimization variables. Fortunately, joint PBA can be handled by primal-dual decomposition and projected subgradient method [16].

IV. NUMERICAL RESULTS AND ANALYSIS

For the simulations, average MBS cell size is arranged to be $500m \times 500m$. QoS requirements of users are uniformly distributed with a mean of 1.5 Mbps. Unless it is stated explicitly otherwise, we use the default simulation parameters given in Table I.

TABLE I: Table of Parameters

Par.	Value	Par.	Value	Par.	Value
η_c^u	3.76	σ_c^u	10 dB	β	0.3
N_0	-174 dBm	ϵ_u	10^{-5}	p_{Δ}	-100 dBm
$\mathbb{E}\{ H_c^u ^2\}$	1	P_m	46 dBm	P_s	30 dBm
B	20 MHz	λ_s	10	λ_u	100

Fig. 3 shows the impacts of receiver sensitivity and residual interference on the normalized sumrate performance for a cluster of 10 users with 1 MHz bandwidth. While the cluster sumrate significantly degrades as ϵ and p_{Δ} increases, the most

dramatic change is observed at $(\epsilon > 10^{-5}, p_{\Delta} > -70 \text{ dBm})$ where NOMA delivers a poor performance with respect to OMA which allocate each user with 100 kHz bandwidth.

Fig. 3: SIC imperfections

Fig. 4 shows the impacts of offloading factor β and cluster size (i.e., number of ICs) L on network sumrate, max-min fairness, and energy & spectrum efficiency ($a = b$) in subfigures Fig. 4a, Fig. 4b, and Fig. 4c, respectively. Similarly, subfigures of Fig. 5 illustrate the effects of UE density λ_u . While subfigures a-b share a common legend, energy and spectrum cost in subfigure d is differentiated by employing solid and dashed lines, respectively. Moreover, we prefer to present results in a normalized scale for a better comparison of different parameters within the same plot.

In Fig. 4a, increasing L (i.e., increasing cluster size or number of ICs) improves the network sumrate which is obviously become more gradual as L gets larger. Moreover, NOMA gain is higher for smaller β values as offloading more users to SBSs give a better chance to form clusters with a higher channel gain disparity. Finally, sumrate performance first increase and then decrease as β increases, which is resulted from the variation in users' SINR due to the user association. In Fig. 4b, max-min fairness decreases as L ranges from 1 to 5. This is intuitive since max-min fairness tends to equalize user rates which reduce if more users share the same bandwidth. Interestingly, $L = 1$ always give a better max-min rate than OMA scheme while $L > 1$ cases do not always provide a better rate for different β values. Please note that max-min fairness curves follow a similar shape of sumrate with respect to β because of the variation in users' SINR due to the user association. Energy efficiency in Fig. 4c, on the other hand, decreases as β increases since more users are associated with the far away MBS. Please also note that putting more users in the same cluster also increases the energy cost as cluster members need to have a certain reception power disparity following from PDSCs. However, spectrum costs plotted in dashed lines increase to compensate the loss in energy efficiency. That is, spectrum and energy costs variate in opposite directions to achieve QoS requirements of UEs.

In Fig. 5, increasing user density increases number of clusters and decreases power and spectrum availability per cluster. Accordingly, we observe that sumrate increases until the network is overloaded (around $\lambda_u = 150$), then it starts decreasing with increasing λ_u . λ_u and number of clusters increase proportionally which decreases the available spectrum

Fig. 4: β vs. L : a) Sumrate, b) Max-Min fairness, and c) Energy & spectrum cost.

Fig. 5: λ_u vs. L : a) Sumrate, b) Max-Min fairness, and c) Energy & spectrum cost.

per cluster. Hence, max-min rate constantly decreases since max-min fairness tends to bring user rates as close as possible. Constant increase in energy and spectrum costs in Fig. 5c is expected as more power and spectrum is needed for more users.

V. CONCLUSION

In this paper, we investigated CF and Joint PBA for NOMA in D2D-Enabled HetNets. Based on a generic imperfect SIC receiver model, we demonstrated the impacts of PDSCs and residual interference on NOMA performance. Based on a multi-partite matching base polynomial time clusterf formation method, PBA problem of HetNets is put into a convex form using GP for different objectives including sumrate maximization, max-min fairness, and energy and/or spectrum cost minimization. Finally, effect of cluster size on NOMA performance is shown for different traffic offloading factors and user density values.

REFERENCES

- [1] H. Elsayw, E. Hossain, and D. I. Kim, "Hetnets with cognitive small cells: user offloading and distributed channel access techniques," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 28–36, Jun. 2013.
- [2] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5g systems with randomly deployed users," *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1501–1505, Dec. 2014.
- [3] 3GPP, *Study on downlink multiuser superposition transmission (MUST) for LTE (Release 13)*. 3rd Generation Partnership Project (3GPP), Dec. 2015, no. TR 36.859.
- [4] S. Timotheou and I. Krikidis, "Fairness for non-orthogonal multiple access in 5g systems," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1647–1651, Oct. 2015.
- [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] A. Celik, R. M. Radaydeh, F. S. Al-Qahtani, A. H. Abd Al-Malek, and M.-S. Alouini, "Resource allocation and cluster formation for imperfect noma in dl/ul decoupled hetnets." [Online]. Available: <http://repository.kaust.edu.sa/kaust/>
- [7] 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.
- [8] M. F. Hanif, Z. Ding, T. Ratnarajah, and G. K. Karagiannis, "A minorization-maximization method for optimizing sum rate in the downlink of non-orthogonal multiple access systems," *IEEE Trans. Signal Process.*, vol. 64, no. 1, pp. 76–88, Jan. 2016.
- [9] Q. Sun, S. Han, C. L. I, and Z. Pan, "On the ergodic capacity of mimo noma systems," *IEEE Wireless Commun. Letters*, vol. 4, no. 4, pp. 405–408, Aug. 2015.
- [10] 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.
- [11] 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.
- [12] R. Jonker and A. Volgenant, "A shortest augmenting path algorithm for dense and sparse linear assignment problems," *Computing*, vol. 38, no. 4, pp. 325–340, 1987.
- [13] S. Boyd, S.-J. Kim, L. Vandenberghe, and A. Hassibi, "A tutorial on geometric programming," *Optimization and Engineering*, vol. 8, no. 1, p. 67, 2007.
- [14] M. Chiang, "Geometric programming for communication systems," *Foundations and Trends in Communications and Information Theory*, vol. 2, no. 12, pp. 1–154, 2005.
- [15] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.
- [16] D. P. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1439–1451, Aug. 2006.