Resource Allocation for Non-Orthogonal Multiple Access in Heterogeneous Networks

Jingjing Zhao*, Yuanwei Liu†, Kok Keong Chai*, Arumugam Nallanathan†, Yue Chen*, and Zhu Han‡
* Queen Mary University of London, London, UK
† King's College London, London, UK
‡ University of Houston, Houton, Tx, USA

***Abstract*—In this paper, novel resource allocation design is investigated for NOMA-enhanced heterogeneous networks (HetNets), where small cell base stations (SBSs) are enabled to communicate with multiple small cell users (SCUs) via the NOMA protocol. The resource allocation problem with the aim of maximizing the sum rate of SCUs is formulated as a many-to-one matching game. Due to the existence of co-channel interference, this game is shown to belong to a class of matching games with peer effects. To solve this game, we propose a novel distributed algorithm where the SBSs and resource blocks (RBs) can interact to decide their desired allocation. The proposed algorithm is proved to converge to a two-sided exchange-stable matching with much lower complexity compared to the centralized method. Simulation results unveil that: 1) The proposed algorithm closely approaches the global optimal solution by around 92.5% within a limited number of iterations; and 2) The developed NOMA-enhanced HetNets scheme achieves a higher sum rate of SCUs compared to the traditional OMA-based HetNets scheme.**

I. Introduction

To meet the surging traffic demands for wireless services and the need for high data rates, cellular networks are trending strongly towards heterogeneity of cells with different transmit power, coverage range and cost of deployment [1–3]. Heterogeneous networks (HetNets) is capable of achieving more spectrum-efficient communications by deploying small cells, e.g., picocells and femtocells, underlaid on the macrocells. Since the spectrum sharing among multi-tier cells can cause both co-tier and cross-tier interference, efficient resource allocation and interference management are the fundamental research challenges for HetNets. In [4], a unified static framework was employed to study the interplay of user association and resource allocation in heterogeneous cellular networks. A novel solution that jointly associated the users to the access points (APs), and allocated the femtocell access points (FAPs) to the service providers (SPs) in an uplink OFDMA network was studied in [5], with the aim of maximizing the total satisfaction of users. Considering the device-to-device (D2D)-enabled multi-tier scenario, a polynomial time-complexity distributed solution approach for the heterogeneous cellular mobile communication systems was presented in [6].

Recently, the non-orthogonal multiple access (NOMA) technique has attracted significant research interests for its potential to enhance spectrum efficiency by allowing multiple users simltaneous transmission in the same resource block (RB) [7]. More specifically, the fundamental concept of NOMA is to facilitate the access of multiple users in an extra dimension—power domain, via different power levels, which is different from conventional orthogonal multiple access (OMA) techniques. Some initial research contributions have been done in terms of NOMA both in single-carrier [8–10] and multi-carrier scenarios [11, 12]. Regarding single carrier NOMA systems, considering users were randomly deployed in a disc, the performances of both outage probability and ergodic rate were investigated in [8]. On the standpoint of energy aspects, a new cooperative NOMA transmission protocol was proposed in [9], in which near NOMA users were regarded as energy harvesting user relays for forwarding messages to far NOMA users. It is worth noting that besides power domain multiplexing ability brought by NOMA, multi-carrier systems are capable of providing additional degrees of freedom offered by multiuser diversity, which motivates researchers to work on multi-carrier NOMA systems. In [11], the authors jointly investigated the power and subcarrier allocation problem in multi-carrier NOMA systems, where the BS worked in a full-duplex mode. Both the optimal performance with applying a monotonic optimization approach and the suboptimal performance with applying a low complexity iterative approach were demonstrated. For maximizing the energy efficiency of the downlink multi-carrier NOMA systems, a low complexity suboptimal subchannel assignment and power proportional factors determination algorithm was proposed in [12], by assuming only two users can be assigned in the same channel.

Despite the fact that there are ongoing research efforts to address the resource allocation problems for both HetNets and NOMA, the solutions for the resource allocation problems in NOMA-enhanced HetNets have not been comprehensively studied in the literature. Note that NOMA-enhanced HetNets pose additional challenges since they bring more co-channel interference to the existing networks. As such, novel resource allocation design for intelligently managing and coordinating various types of interference are more than desired, which motivates us to develop this work. We focus on the study of resource allocation for NOMA-enhanced HetNets with the aim of maximizing the sum rate of small cell users (SCUs). Particularly, we consider the downlink NOMA-enhanced HetNets scenario, where each small base station (SBS) communicates with multiple SCUs via the NOMA protocol. We also allow multiple SBSs to reuse the same RB occupied by a macro cell user (MCU) to further improve the resource utilization. Since the centralized approach is with high complexity to

978-1-4673-8999-0/17/$31.00 ©2017 IEEE

solve the formulated optimization problem due to the existence of co-channel interference, the matching theory is adopted to develop a distributed resource allocation algorithm. The primary contributions of this paper can be summarized as follows. Considering the NOMA-enhanced HetNets system, we propose a distributed resource allocation algorithm for SBSs based on the matching theory to maximize the sum rate of SCUs. We mathematically prove that the complexity of the proposed algorithm is lower than the centralized method. Simulation results demonstrate that the proposed algorithm achieves near-optimal performance and significantly outperforms the conventional OMA-based HetNets.

II. Network Model

A. System Description

Consider a downlink K-tier HetNets model, where the first tier represents a single macro cell and the other tiers represent the small cells such as pico cells and femto cells. The macro base station (MBS) provides the basic coverage, and SBSs are deployed in the coverage area of the MBS to enhance capability. The set of SBSs is represented by $\mathcal{B}^s = \{1, ..., B\}$. The MBS serves a set of M macro cellular users (MCUs), i.e., $U_m = \{1, ..., M\}$. There are M RBs, and each MCU occupies a RB. For the sake of simplicity, we use the same index of the RBs as the MCUs, and thus the set of RBs is represented by $\mathcal{RB} = \{1, ..., M\}$. The macro cell and small cells reuse the same set of RBs, and thus we refer to the small cells as the underlay tier. Each SBS b selects one RB from the available RBs, and serves multiple SCUs, i.e., $U_b = \{1, ..., K\}$, with applying NOMA techniques. The detailed cellular layout is shown in Fig. 1.

In this work, we allow multiple SBSs to reuse the same RB to improve the spectrum efficiency. The maximum number of SBSs occupying the same RB is restricted to q_{max}. We assume that the total number of SBSs allowed for spectrum access, i.e., $M \times q_{max}$, is larger than B. Thus, all the SBSs can be served. Since the spectrum sharing brings in both co-tier and cross-tier interference, efficient resource allocation is needed for the NOMA-enhanced HetNets system. In our work, we assume that the user association to the MBS and SBSs are completed prior to the resource allocation.

B. Channel Model

The NOMA based transmission requires to apply the superposition coding (SC) technique at the SBSs and successive interference cancellation (SIC) technique at the SCUs. The vector $\mathbf{a}_b \in \mathbb{R}^{1\times K}$ with the elements $a_{b,k}$ represents the power allocation coefficients for the SCUs in each small cell. We assume that the power allocation coefficients are fixed. The SBS b sends K messages to the destinations on RB m, based on the NOMA principle, i.e., b sends $\sum_{k=1}^{K} a_{b,k}^n x_{b,k}^n$, where $x_{b,k}^n$ is the message for $SCU_{b,k}$. The received signal at the k-th SCU, i.e., $k \in \{1, ..., K\}$, served by the b-th SBS, i.e., $b \in \{1, ..., B\}$, on the m-th RB is given by

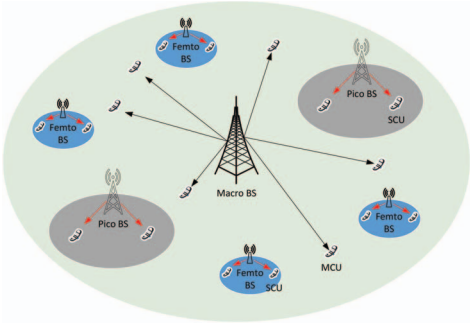

Fig. 1: Illustration of NOMA-enhanced HetNets.

$$y_{b,k}^n = \underbrace{f_{b,k}^m \sqrt{p_b a_{b,k}} x_{b,k}^m}_{\text{desired signal}} + \underbrace{f_{b,k}^m \sum_{k'=k}^{K} \sqrt{p_b a_{b,k'}} x_{b,k'}^m}_{\text{interference from NOMA users}} + \underbrace{\zeta_{b,k}^m}_{\text{noise}} + \underbrace{\sum_{m=1}^{M} \lambda_{m,b} h_{m,b,k} \sqrt{p_m} x_m}_{\text{cross-tier interference}} + \underbrace{\sum_{b*\neq b} \lambda_{b*,b} g_{b*,b,k}^m \sqrt{p_{b^*}} x_{b*}^m}_{\text{co-tier interference}}, \quad (1)$$

where $x_{b,k}^m$, x_m are the symbols transmitted from the b-th SBS to its serving SCU k, i.e, $SCU_{b,k}$, and from the MBS to the MCU m, respectively. $f_{b,k}^m$, $h_{m,b,k}$, and $g_{b*,b,k}^m$ are the channel coefficients between SBS b and $SCU_{b,k}$, that between the MBS and $SCU_{b,k}$, and that between SBS b^* and $SCU_{b,k}$ on RB m, respectively. p_b and p_m are the total transmit power of the SBS b and the transmit power from the MBS to the MCU m. respectively. $\lambda_{m,b}$ represents the RB allocation indicator for SBSs, i.e., if SBS b occupies RB m, $\lambda_{m,b} = 1$; otherwise, $\lambda_{m,b} = 0$. $\lambda_{b*,b}$ represents the presence of co-tier interference, i.e., if the SBS b and b^* reuse the same RB, $\lambda_{b*,b} = 1$; otherwise, $\lambda_{b*,b} = 0$. $\zeta_{b,k}^n$ is the additive white gaussian noise (AWGN) at $SCU_{b,k}$ with variance σ^2.

NOMA systems exploit the power domain for multiple access, where different users are served at different power levels. For illustration, we assume that the SIC decoding order is as the index order of the SCUs served by each SBS, i.e., the k-th SCU can always decode the signals of the $\{1, ..., (k-1)\}$-th SCUs. Specifically, the k-th SCU first successively subtracts the messages of the SCUs $j < k$, and then obtains its own information by regarding the messages of the SCUs $i > k$ as noise. Therefore, according to the received signal expressed in (1), the received SINR at the k-th SCU served by the b-th SBS on RB m to decode its own information is given by

$$\gamma_{b,k,k}^m = \frac{\left|f_{b,k}^m\right|^2 p_b a_{b,k}^m}{I_N^{k,k} + I_{co}^k + I_{cr}^k + \sigma^2}, \quad (2)$$

where $I_N^{k,k} = |f_{b,k}^m|^2 p_b \sum_{i=k+1}^{K} a_{b,i}^m$ is the interference from the NOMA users served by the same SBS, $I_{co}^k = \sum_{b*\neq b} \lambda_{b^*,b} p_{b^*} \left|g_{b*,b,k}^m\right|^2$ is the co-tier interference from the other SBSs reusing the same RB, and $I_{cr}^k = \sum_m \lambda_{m,b} p_m |h_{m,b,k}|^2$ is the cross-tier interference from the MBS. Here, $|f_{b,k}^m|^2 = |\hat{f}_{b,k}^m|^2 (d_{b,k})^{-\alpha}$, $|g_{b*,b,k}^m|^2 =$

$|\hat{g}^m_{b*,b,k}|^2(d_{b*,b,k})^{-\alpha}$, and $|h_{m,b,k}|^2 = |\hat{h}_{m,b,k}|^2(d_{m,b,k})^{-\alpha}$. $\hat{f}^m_{b,k}$, $\hat{g}^m_{b^*,b,k}$ and $\hat{h}_{m,b,k}$ are small-scale fading with $\hat{f}^m_{b,k} \sim \mathcal{CN}(0,1)$, $\hat{g}^m_{b^*,b,k} \sim \mathcal{CN}(0,1)$ and $\hat{h}_{m,b,k} \sim \mathcal{CN}(0,1)$. $d_{b,k}$ is the distance from SBS b to $SCU_{b,k}$. $d_{b*,b,k}$ is the distance from SBS b^* to $SCU_{b,k}$, and $d_{m,b,k}$ is the distance from the MBS to $SCU_{b,k}$.

Note that $SCU_{b,K}$ can decode the signals of all the other SCUs served by SBS b, thus the SINR received at $SCU_{b,K}$ is expressed as

$$\gamma^m_{b,K,K} = \frac{\left|f^m_{b,K}\right|^2 p_b a^m_{b,K}}{I^K_{co} + I^K_{cr} + \sigma^2}. \tag{3}$$

For the SCUs served by the same SBS, the received SINR for SCU k for the SCU j's required signal is given by

$$\gamma^m_{b,k,j} = \frac{\left|f^m_{b,k}\right|^2 p_b a^m_{b,j}}{I^{k,j}_N + I^k_{co} + I^k_{cr} + \sigma^2}, \tag{4}$$

where $I^{k,j}_N = \left|f^m_{b,k}\right|^2 p_b \sum_{i=j+1}^{K} a^m_{b,i}$. The interference cancellation is successful if the SCU k's received SINR for the SCU j's signal is larger or equal to the received SINR of SCU j for its own signal [8, 11]. Therefore, the condition of our given SIC decoding order is given by

$$\frac{\left|f^m_{b,k}\right|^2 p_b a^m_{b,j}}{I^{k,j}_N + I^k_{co} + I^k_{cr} + \sigma^2} \geq \frac{\left|f^m_{b,j}\right|^2 p_b a^m_{b,j}}{I^{j,j}_N + I^j_{co} + I^j_{cr} + \sigma^2}. \tag{5}$$

To guarantee the service qualities of the MCUs, we give an interference threshold I_{thr} to the aggregated interference caused to the MCUs from the links in the underlay tier. The aggregated interference experienced on the MCU m is given by

$$I_m = \sum_{b=1}^{B} \lambda_{m,b} p_b \left|t_{b,m}\right|^2, \tag{6}$$

where $|t_{b,m}|^2 = \left|\hat{t}_{b,m}\right|^2 (d_{b,m})^{-\alpha}$, and $\hat{t}_{b,m}$ is small-scale fading with $\hat{t}_{b,m} \sim \mathcal{CN}(0,1)$. $d_{b,m}$ is the distance from SBS b to MCU m.

III. Problem Formulation and Proposed Optimization Method

A. Optimization Problem Formulation

Based on the SINR expressions of $SCU_{b,k}$, $\forall b \in \{1,...,B\}, k \in \{1,...,K\}$ in (2) and (3), the data rate of $SCU_{b,k}$ over the RB m can be calculated as

$$R^m_{b,k} = \lambda_{m,b} \log_2 \left(1 + \frac{\left|f^m_{b,k}\right|^2 p_b a^m_{b,k}}{I^{k,k}_N + I^k_{co} + I^k_{cr} + \sigma^2}\right). \tag{7}$$

For facilitating the presentation, we denote $\lambda \in \mathbb{R}^{M \times B}$, $\mathbf{a} \in \mathbb{R}^{B \times K}$ as the collections of optimization variables $\lambda_{m,b}$, and $a_{b,k}$, respectively. The system objective is to maximize the SCUs' sum data rate with the interference constraints of the MCUs satisfied, which can be expressed as follows:

$$\max_{\lambda} \sum_{b=1}^{B} \sum_{k=1}^{K} \sum_{m=1}^{M} \hat{R}^m_{b,k}(\lambda), \tag{8a}$$

$$s.t. \quad \lambda_{m,b} \in \{0,1\}, \quad \forall m, b, \tag{8b}$$

$$\sum_{m} \lambda_{m,b} \leq 1, \quad \forall b, \tag{8c}$$

$$\sum_{b} \lambda_{m,b} \leq q_{max}, \quad \forall m, \tag{8d}$$

$$I_m \leq I_{thr}, \forall m. \tag{8e}$$

Constraint (8b) and (8c) are imposed to guarantee that each SBS occupies no more than one RB. Constraint (8d) limits the maximum number of SBSs, i.e., q_{max}, reusing each RB. Constraint (8e) guarantees that the total received interference at each MCU is less or equal to threshold I_{thr}.

To find the global optimal solution of (8), we need to fully search for all the possible combinations of scheduling RBs to SBSs. Thus, even for a centralized algorithm, it is not feasible in practical systems to solve this problem optimally. However, since λ is a binary variable, we can formulate the RB allocation as a many-to-one matching problem.

B. Many-to-One Matching Game

To proceed with formulating the matching problem, we first introduce some notations.

Definition 1. *In the many-to-one matching model, a* matching Φ *is a function from the set* $\mathcal{RB} \cup \mathcal{B}^s$ *into the set of all subsets of* $\mathcal{RB} \cup \mathcal{B}^s$ *such that 1)* $|\Phi(b)| = 1, \forall b \in \mathcal{B}^s$*; 2)* $|\Phi(m)| \leq q_{max}$*,* $\forall m \in \mathcal{RB}$*; 3)* $\Phi(b) = m$ *if and only if* $b \in \Phi(m)$.

For the conditions in the definition, condition (1) implies that each SBS can only be matched with one RB; condition (2) gives the quota q_{max} of the maximum number of SBSs that can be matched to each RB; and condition (3) implies that if SBS b is matched with RB m, then RB m is also matched with SBS b.

The utility of SBS b is defined as the sum rate of all the serving SCUs minus its cost for occupying RB m, which is given by

$$U_b = \sum_{k=1}^{K} R^m_{b,k} - \beta p_b \left|g_{b,m}\right|^2, \tag{9}$$

where $\beta \in \mathbb{R}^+$ is the fixed coefficient with unit interference of SBS b bringing to the m-th MCU.

The utility of RB m is defined as the sum rate of the occupying SCUs plus the fees it charges the SBSs for causing interference to the corresponding MCU, and thus the utility function of RB m can be expressed as

$$U_m = \sum_{b=1}^{B} \lambda_{m,b} \left(\sum_{k=1}^{K} R^m_{b,k} + \beta p_b \left|g_{b,m}\right|^2\right), \tag{10}$$

To start the matching process, both SBSs and RBs need to set up the preference lists of their own. It is a descending order list formed by each side of the players according to their preference to the other side of the players. For each SBS $b \in \mathcal{B}^s$, it forms a descending order preference list $\mathcal{BLIST}_b$ according to its utilities over all the RBs, i.e., the first RB in the preference list is the RB m leading to the maximum U_b^m. Similarly, each RB $m \in \mathcal{RB}$ forms a preference list $\mathcal{RBLIST}_m$ over all the SBSs with the descending order of its own utilities.

Because of the existence of the co-tier interference among SBSs reusing the same RB, the utility of each SBS depends not only on the RB it is matched with, but also on which other SBSs are matched to the same RB. Thus, we say the proposed matching game is with *peer effects*. This is different from the conventional matching games where players have fixed preference lists [13, 14]. There is a growing literature studying many-to-one matchings with peer effects [15, 16]. However, these researches find that designing matching mechanisms is significantly more challenging when peer effects are considered. Motivated by the housing assignment problem in [17], we propose an extended matching algorithm for the many-to-one matching problem with peer effects in the following.

Different from the traditional Gale Shapley (GS) Algorithm [18], the *swap operations* between any two SBSs to exchange their matched RBs is enabled. To better describe the interdependencies between the players' preferences, we first define the concept of *swap matching* as follows:

$$\begin{aligned}\Phi_b^{b'} =& \{\Phi \setminus \{(b, \Phi(b)), (b', \Phi(b'))\}\} \cup \\ & \{(b, \Phi(b')), (b', \Phi(b))\}, \end{aligned} \tag{11}$$

where SBSs b and b' switch places while keeping other SBSs and RBs' matchings unchanged.

Based on the concept of *swap operations*, the *swap-blocking* pair is defined as follows:

Definition 2. *A pair of SBSs (b, b') is a* swap-blocking *pair if and only if*

1) $\forall s \in \{b, b', \Phi(b), \Phi(b')\}, U_s(\Phi_b^{b'}) \geq U_s(\Phi)$ and;

2) $\exists s \in \{b, b', \Phi(b), \Phi(b')\}$, such that $U_s(\Phi_b^{b'}) > U_s(\Phi)$, where $U_s(\Phi)$ represents the utility of the player s under the matching state Φ.

Note that the above definition implies that if two SBSs want to switch between two RBs, the RBs involved must approve the swap. Condition 1) implies that the utilities of all the involved players should not be reduced after the swap operation. Condition 2) indicates that at least one of the players' utilities is increased after the swap operation. This avoids looping between equivalent matchings where the utilities of all involved agents are indifferent.

C. Proposed Resource Allocation Algorithm

In this subsection, we first propose an initialization algorithm (IA) based on the GS algorithm to obtain the initial matching state, which is inspired by the work in [19]. After the initialization, we proceed with swap operations among SBSs to further improve the performance.

1) Initialization Algorithm: In the initialization algorithm, SBSs and RBs first initialize their own preference lists. The list of all the SBSs that are not matched with any RB is denoted by $\mathcal{UNMATCH}$. In the matching process, each SBS proposes to its most preferred RB, then each RB accepts the most preferred SBS and rejects the others. This process continues until the set $\mathcal{UNMATCH}$ goes empty. The details of the initialization algorithm are as shown in **Algorithm 1**.

Algorithm 1 Initialization Algorithm (IA)

1: Construct the preference lists of the SBSs $\mathcal{BLIST}_b, b \in \mathcal{B}^s$;
2: Construct the preference lists of the RBs $\mathcal{RBLIST}_m, m \in \mathcal{RB}$;
3: Construct the set of the SBSs that are not matched $\mathcal{UNMATCH}$;
4: **while** $\mathcal{UNMATCH} \neq \emptyset$ **do**
5: **for** $\forall b \in \mathcal{UNMATCH}$ **do**
6: SBS b proposes to its most preferred RB that has never rejected it before;
7: **end for**
8: **for** $\forall m \in \mathcal{RB}$ **do**
9: **if** $\sum_{b \in \mathcal{B}^s} \eta_{m,b} \leq q_{max}$ **then**
10: RB m keeps all the proposed SBSs;
11: Remove the matched SBSs from $\mathcal{UNMATCH}$;
12: **else**
13: RB m keeps the most preferred q_{max} SBSs, and rejects the others;
14: Remove the matched SBSs from $\mathcal{UNMATCH}$;
15: Keep the rejected SBSs in $\mathcal{UNMATCH}$.
16: **end if**
17: **end for**
18: **end while**

2) Swap Operations Enabled Matching Algorithm: After the initialization of the matching state based on the IA, swap operations among SBSs are enabled to further improve the performance of the resource allocation algorithm. The details of the proposed swap operations enabled matching algorithm (SOEMA) are shown in **Algorithm 2**. The SOEMA is composed of three steps. Step 1 is to do the initialization based on the IA. Step 2 focuses on the the swap operations between the SBSs. Each SBS keeps searching for all the other SBSs to check whether there exists a swap-blocking pair. The swap-matching process continues until there exists no swap-blocking pair, and then the algorithm goes to step 3, i.e., the end of the algorithm. Note that to prevent SBS b looping in the swap operations with another SBS b', we set the flag $\mathcal{SR}_{b,b'}$ to record the time that SBS b and b' swap their allocated RBs. Each SBS b can at most swap with another SBS b' twice.

3) Property Analysis: As stated in [18], there is no longer a guarantee that a traditional "pairwise-stability" exists when players care about more than their own matching, and, if a stable matching does exist, it can be computationally difficult to find. The authors in [17] focused on the *two-sided exchange-stable matchings*, which is defined as follows:

Algorithm 2 Swap Operations Enabled Matching Algorithm (SOEMA)

– ***Step 1: Initialization***
Matching by the Initialization Algorithm (IA);
Obtain the initial matching state: Φ_0;
Initialize the number of swapping requests that SBS b sends to b', i.e., $\mathcal{SR}_{b,b'} = 0$;
– ***Step 2: Swap-matching process:***
For each SBS b, it searches for another SBS b' to form a swap-blocking pair;
if (b, b') forms a swap-blocking pair along with $m = \Phi(b)$, and $m' = \Phi(b')$, as well as $\mathcal{SR}_{b,b'} + \mathcal{SR}_{b',b} < 2$ **then**
  Update the current matching state to $\Phi_b^{b'}$;
  $\mathcal{SR}_{b,b'} = \mathcal{SR}_{b,b'} + 1$;
else
  Keep the current matching state;
end if
Repeat ***Step 2*** until there is no swap-blocking pair.
– ***Step 3: End of the algorithm***

Definition 3. *A matching* Φ *is* two-sided exchange-stable *if there does not exist a* swap-blocking *pair.*

The *two-sided exchange stability* is a distinct notion of stability compared to the traditional notion of stability of [18], but one that is relevant to our situation where agents can compare notes with each other.

Proposition 1. *The final matching* Φ^* *of SOEMA is a* two-sided exchange-stable *matching.*

Proof. Assume that there exists a swap-blocking pair (b, b') in the final matching Φ^* satisfying that $\forall s \in \{b, b', \Phi(b), \Phi(b')\}, U_s\left((\Phi^*)_b^{b'}\right) \geq U_s(\Phi^*)$ and $\exists s \in \{b, b', \Phi(b), \Phi(b')\}$, such that $U_s\left((\Phi^*)_b^{b'}\right) > U_s(\Phi^*)$. According to SOEMA, the algorithm does not terminate until all the swap-blocking pairs are eliminated. In other words, Φ^* is not the final matching, which causes conflict. Therefore, there does not exist a swap-blocking pair in the final matching, and thus we can conclude that the proposed algorithm reaches a two-sided exchange stability in the end of the algorithm. □

The complexity of SOEMA is composed of two main parts, i.e., the IA and the swap-matching phases. For the IA, the complexity of setting up the preference lists of SBSs and RBs is $\mathcal{O}(BM^2)$. For the swap-matching phase, since we restrict that each SBS b can at most swap its allocated RB with another SBS b' twice, the number of potential swap operations is upper bounded by $2 \times \binom{B}{2}$, which is with the complexity of $\mathcal{O}(B^2)$. In contrast, the complexity of the centralized optimal method increases exponentially with B and M. Therefore, both IA and SOEMA have significantly lower complexities than the centralized optimal method.

IV. Numerical Results

In this section, we investigate the performance of the proposed resource allocation algorithm through simulations. The adopted simulation parameters are given in Table I. We assume that each SBS serves at most two SCUs simultaneously via the NOMA protocol, to limit the co-channel interference and to lower the hardware complexity and processing delay[1]. The optimal performance obtained by the centralized method is given as the baseline. We compare SOEMA with IA in the NOMA-HetNets system to show differences among their performances. In addition, we also consider the performance of the traditional OMA-HetNets system where each SBS communicates with at most one SCU in a transmission interval. In order to have a fair comparison, the resource allocation result for the OMA-based HetNets is also obtained by utilizing SOEMA (SOEMA-OMA) and IA (IA-OMA), respectively.

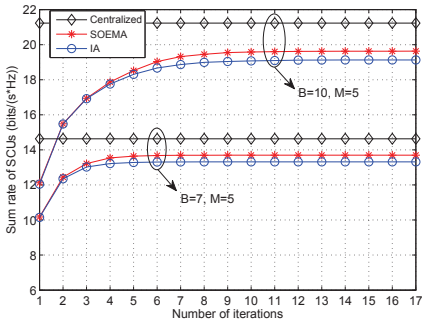

Fig. 2: Convergence of the proposed algorithms for different number of RBs and SBSs.

TABLE I: Parameter Values Used in Simulations

Macro cell radius, small cell radius	300 m, 30 m
Transmit power of MBS and SBSs	43 dBm, 23dBm
Noise power spectral density	-174 dBm/Hz
Interference threshold at each MCU	-70dBm
Maximum number of SBSs reusing each RB	2

Fig. 2 illustrates the convergence of the proposed algorithms, i.e., IA and SOEMA, with different numbers of RBs M and SBSs B. It can be seen that IA and SOEMA both converge within a small number of iterations for different values of M and B. Besides, both IA and SOEMA need more iterations to converge with a larger number of RBs and SBSs. For example, when $B = 7, M = 5$, SOEMA and IA converge in less than 6 iterations on average. When $B = 10, M = 5$, SOEMA and IA converge to a stationary point at around 12 iterations. This is due to the fact that additional players participating in the matching game results in additional searching dimensions in the possible matching solutions. It is also shown in Fig. 2 that the proposed algorithm performs very close to the centralized method. In particular, for the case of $B = 10, M = 5$, SOEMA gets around 92.5% of the sum rate of SCUs achieved by the centralized method.

Fig. 3 plots the sum rate of SCUs versus different numbers of SBSs B in the network, for number of RBs $M = 10$. As can be seen from Fig. 3 that the sum rate increases monotonically

[1]NOMA requires SIC at the receivers. A user performing SIC needs to demodulate and decode the signals transmitted to other receivers. Therefore, the hardware complexity and processing delay increases with the number of users multiplexed on the same RB.

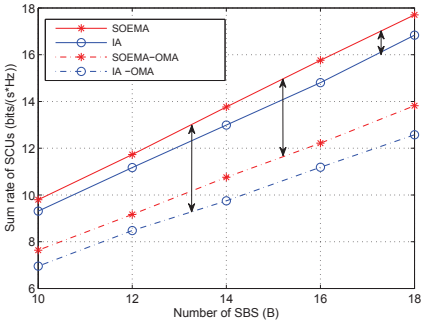

Fig. 3: Sum rate of the SCUs with different number of small cells, with $M = 10$.

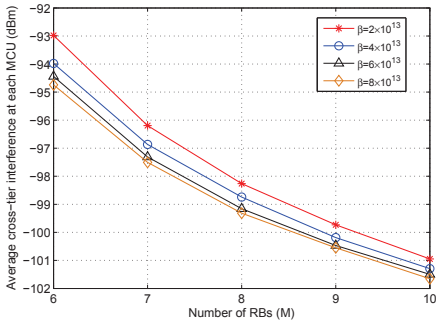

Fig. 4: Average received cross-tier interference at each MCU, with $B = 12$.

with the number of SBSs due to the exploitation of multi-user diversity gain and the assumption that the total number of SBSs allowed for spectrum access is always larger than the number of SBSs to be scheduled. It is also observed that SOEMA achieves a higher sum rate compared to the IA due to the involvement of the swap operations between the potential swap-blocking pairs. Besides, we also notice that the OMA-HetNets system achieves a lower sum rate compared to the NOMA-HetNets system since OMA leads to underutilized spectrum resource. In particular, at the point of $B = 18, M = 10$, SOEMA achieves roughly a 10%, 30% and 45% higher sum rate than IA, SOEMA-OMA, and IA-OMA, respectively.

Fig. 4 shows the impact of β on the average received interference at each MCU. It is demonstrated that the average received interference at each MCU decreases with the number of MCUs. We also note that the average interference decreases with the larger value of β. Recall (9) and (10), and we observe that more importance is given to the received interference at MCUs rather than the rate of SCUs with larger value of β, which leads to the performance in Fig. 4.

V. Conclusions

In this paper, the resource allocation problem in NOMA-enhanced HetNets was studied. Aiming at maximizing the system sum rate, the resource allocation optimization problem was modeled as a many-to-one matching game, and a novel distributed algorithm was proposed to solve the optimization problem efficiently. It was proved mathematically that the proposed algorithm was stable and with lower complexity compared to the centralized method. Simulation results demonstrated that the proposed algorithm achieved the near-optimal sum rate of SCUs and obtained significant improvements compared to the conventional OMA-based HetNets scheme.

VI. Acknowledgement

This work was partially supported by the US NSF under grant number CNS-1646607, ECCS-1547201, CCF-1456921, CNS-1443917 and ECCS-1405121.

References

[1] H. S. Dhillon, R. K. Ganti, F. Baccelli, and J. G. Andrews, "Modeling and analysis of k-tier downlink heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 550–560, Apr. 2012.

[2] X. Lagrange, "Multitier cell design," *IEEE Commun. Mag.*, vol. 35, no. 8, pp. 60–64, Aug. 1997.

[3] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, Jun. 2013.

[4] D. Fooladivanda and C. Rosenberg, "Joint resource allocation and user association for heterogeneous wireless cellular networks."

[5] S. Bayat, R. H. Louie, Z. Han, B. Vucetic, and Y. Li, "Distributed user association and femtocell allocation in heterogeneous wireless networks," *IEEE Trans. Commun.*, vol. 62, no. 8, pp. 3027–3043, Jul. 2014.

[6] M. Hasan and E. Hossain, "Distributed resource allocation in D2D-enabled multi-tier cellular networks: An auction approach," in *Proc. of the IEEE Int. Conf. on Commun. (ICC)*, London, Jun. 2015, pp. 2949–2954.

[7] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, C. L. I, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

[8] 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.

[9] Y. Liu, Z. Ding, M. Elkashlan, and H. V. Poor, "Cooperative non-orthogonal multiple access with simultaneous wireless information and power transfer," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 938–953, Apr. 2016.

[10] 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, July 2016.

[11] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal joint power and subcarrier allocation for full-duplex multicarrier non-orthogonal multiple access systems," *submitted to IEEE Trans. Wireless Commun.*, 2016. [Online]. Available: https://arxiv.org/abs/1607.02668

[12] F. Fang, H. Zhang, J. Cheng, and V. C. M. Leung, "Energy-efficient resource allocation for downlink non-orthogonal multiple access network," *IEEE Trans. Commun.*, vol. 64, no. 9, pp. 3722–3732, Sep. 2016.

[13] Y. Gu, Y. Zhang, M. Pan, and Z. Han, "Matching and cheating in device to device communications underlying cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 10, pp. 2156–2166, Oct. 2015.

[14] M. Hasan and E. Hossain, "Distributed resource allocation for relay-aided device-to-device communication: A message passing approach," *IEEE Wireless Commun.*, vol. 13, no. 11, pp. 6326–6341, Jul. 2014.

[15] B. Dutta and J. Massó, "Stability of matchings when individuals have preferences over colleagues," *Journal of Economic Theory*, vol. 75, no. 2, pp. 464–475, 1997.

[16] I. E. Hafalir, "Stability of marriage with externalities," *International Journal of Game Theory*, vol. 37, no. 3, pp. 353–369, Mar. 2008.

[17] E. Bodine-Baron, C. Lee, A. Chong, B. Hassibi, and A. Wierman, "Peer effects and stability in matching markets," in *Algorithmic Game Theory*. Springer, 2011, pp. 117–129.

[18] A. E. Roth and M. A. O. Sotomayor, *Two-sided matching: A study in game-theoretic modeling and analysis*. Cambridge University Press, 1992, no. 18.

[19] Y. Zhang, Y. Gu, M. Pan, and Z. Han, "Distributed matching based spectrum allocation in cognitive radio networks," in *Proc. of the IEEE Global Commun. Conf. (GLOBECOM)*, Austin, Dec. 2014, pp. 864–869.