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/TWC.2017.2664832, IEEE Transactions on Wireless Communications

Joint Uplink Base Station Association and Power Control for Small-Cell Networks with Non-Orthogonal Multiple Access

Li Ping Qian, *Senior Member, IEEE*, Yuan Wu, *Senior Member, IEEE*, Haibo Zhou, *Member, IEEE*, and Xuemin (Sherman) Shen, *Fellow, IEEE*

***Abstract*—Since the non-orthogonal multiple access (NOMA) with successive interference cancellation (SIC) can achieve superior spectral-efficiency and energy-efficiency, the concept of SCN using NOMA with SIC is proposed in this paper. Due to the difference of small-cell base stations' locations, each mobile user perceives different channel gains to different small-cell base stations. Therefore, it is important to associate a mobile user with the right base station and control its transmit power for the uplink SCN using NOMA with SIC. However, the already-challenging base station association problem is further complicated by the need of transmit power control, which is an essential component to manage co-channel interference. Despite its importance, the joint base station association and power control optimization problem that maximizes the system-wide utility and at the same time minimizes the total transmit power consumption for the maximum utility has remained largely unsolved for the uplink SCN using NOMA with SIC, mainly due to its non-convex and combinatorial nature. To solve this problem, we first present a formulation transformation that captures two interactive objectives simultaneously. Then, we propose a novel algorithm to solve the equivalently transformed optimization problem based on the coalition formation game theory and the primal decomposition theory in the framework of simulated annealing. Finally, theoretical analysis and simulation results are provided to demonstrate that the proposed algorithm is guaranteed to converge to the global optimal solution in polynomial time.**

***Index Terms*—Small-cell network, Non-orthogonal multiple access, Power control, Small-cell Base station association, System-wide utility maximization, Power minimization.**

I. Introduction

The rapid growth of wireless broadband services renders radio resources even scarer, especially in areas with heavy user demands [1]–[3]. This drives the research community to design more spectrum-efficient and energy-efficient wireless network systems that cope with the scarcity of radio resources.

Manuscript received August 1, 2016; revised November 3, 2016 and January 1, 2017; accepted January 29, 2017. The associate editor coordinating the review of this paper and approving it for publication was B. Natarajan.

This work was supported in part by the National Natural Science Foundation of China (Project number 61379122), by the Zhejiang Provincial Natural Science Foundation of China (Project number LR16F010003 and LR17F010002), and by the Natural Sciences and Engineering Research Council (NSERC), Canada.(*Corresponding author: Yuan Wu.*)

L. P. Qian and Y. Wu are with College of Information Engineering, Zhejiang University of Technology, Hangzhou, China (email: {lpqian,iewuy}@zjut.edu.cn).

H. Zhou and X. Shen are with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: h53zhou@uwaterloo.ca, xshen@bbcr.uwaterloo.ca).

To this end, two-fold efforts have been spent in the literature. On one hand, the deployment of heterogeneous or small-cell networks (SCNs) has been emerging as a promising solution to improve spectrum and energy efficiency as well as expand indoor and cell-edge coverage [4]–[13]. The SCN concept is to deploy various classes of small-cell/lower-power base stations such as picocells, femtocells, and perhaps relay base stations (BSs) underlaid in a macro-cellular network. On the other hand, the so-called non-orthogonal multiple access (NOMA) with successive interference cancellation (SIC) has been investigated as a potential alternative to orthogonal multi-user access to further improve the spectrum-efficiency of cellular networks [Chapter 6, 11], [15]. Therefore, the implementation of SCN using NOMA with SIC will exhibit promising benefits on spectrum-efficiency and energy-efficiency. However, since the SCN using NOMA with SIC is designed based on the SCN architecture, it also suffers from the severe co-channel interference and the unbalanced traffic load distribution of each cell with massive deployment of small cells. For this, it is significant to efficiently allocate BSs (including small-cell BSs and macrocell BSs) and transmit power to mobile users (MUs) which perceive different channel gains to different BSs in the SCN using NOMA with SIC. In particular, the BS association is to establish spectrum-efficient connections between MUs and BSs according to the locations of MUs and the traffic load distribution in the geometric area. Furthermore, the power control is to ensure that each MU transmits an appropriate amount of power to maintain its communication quality without imposing excessive interference on other MUs.

The conventional BS association was realized according to the signal-to-noise ratio (SNR) [16], [17]. Specifically, each MU is associated with the BS providing the strongest SNR. However, such a max-SNR BS association scheme is potentially unsuitable for SCNs, since it does not account for the traffic load distribution of each cell, hence resulting in severe load imbalance and inefficient spectrum and energy utilization. To cope with this problem, the joint optimization of BS association and power control becomes essential for SCNs because small-cell BSs are often deployed to alleviate traffic "hot-spots" with higher-than-average user density and improve the spectrum and energy efficiency of the network [18]–[28]. However, most previous work studied the joint optimization of BS association and power control for the single objective, such as power minimization [18]–[24] and system-wide utility maximization [25]–[28]. To the best of our knowledge, there is

1536-1276 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

few work taking into account these two interactive objectives simultaneously, although the spectrum-efficiency and energy-efficiency can be further improved by striking a balance between the system-wide utility maximization and power minimization. Furthermore, these works assume either the orthogonal multiple access (e.g., orthogonal frequency division multiple access) or the traditional non-orthogonal multiple access (i.e., pure power control).

In this paper, we study an uplink SCN consisting of multiple macro-cell BSs (MBSs), small-cell BSs (SBSs) and MUs, in which all BSs and MUs are equipped with a single antenna each and all channel state information (CSI) related to MUs is perfect to each BS. Driven by its superior spectrum efficiency, we introduce the NOMA with SIC to the uplink SCN, in which the MUs associated with one BS adopt the simultaneous multi-user transmission via superposition coding, and each BS adopts the SIC receiver to decode the signals from its associated MUs sequentially. Due to the complicated coupling among the signal-to-interference-and-noise ratios of simultaneous transmissions, there exist multiple solutions of optimal power allocation that maximizes the system-wide utility. To improve the spectrum-efficiency and energy-efficiency of the uplink SCN using NOMA with SIC, we study the joint optimization of BS association and power control that first maximizes the system-wide utility while keeping every served MU meeting the minimum average-rate requirement, and further minimizes the total power consumption based on the maximum utility. In particular, the main contributions of this paper are summarized as follows:

- **NOMA with SIC:** We present the NOMA scheme with SIC for the uplink SCN. While the out-of-cell interference has to be considered, an SIC receiver at each BS is used to decode the information of all associated MUs sequentially for mitigating the in-cell interference.
- **Novel problem formulation:** We present the joint optimization of BS association and power control problem on maximizing the system-wide utility and meanwhile minimizing the total transmit power subject to the maximum utility for the uplink SCN using NOMA with SIC. This paper equivalently integrates two interactive objectives into a single-objective optimization problem via weighted summing the system-wide utility and the total transmit power consumption with carefully chosen weights.
- **Efficient algorithm design:** The equivalent single-objective optimization problem is a mixed-integer nonconvex program involving the assignments of MUs to BSs and power allocation, and thus difficult to solve. We therefore propose a novel algorithm, referred to as PCSUM (Power Controlled System-wide Utility Maximization), to efficiently and optimally solve it based on the coalition formation game theory and the primal decomposition theory in the framework of simulated annealing. Our proposed algorithm is guaranteed to converge to the globally optimal solution in polynomial time.

The rest of the paper is organized as follows. Section II summarizes the related work on the joint BS association and power control. Section III introduces the system model about the uplink SCN using NOMA with SIC, and problem formulation on the joint optimization of BS association and power control. Section IV proposes the PCSUM algorithm to optimize the BS association and transmit power levels. Section V evaluates the performance of PCSUM through several simulations. Section VI concludes this paper.

II. Related Work

Driven by the significance of the BS association and power control that can efficiently improve the spectrum and energy efficiency, this topic has attracted lots of research interests in the past two decades [16], [17], [29]–[35]. Specifically, most prior work on the BS association and power control can be traced back to traditional cellular networks (i.e., macrocell-only networks) [16], [17], [29]–[35]. For example, work [16] and [17] decoupled BS association from the optimization of per-link transmit power through associating a MU with the BS providing the strongest signal. Work [30]–[35] took into account the BS association under the assumption that transmit power is fixed. Although the BS association was investigated in these earlier work, it fails to characterize how transmit power is efficiently used to transmit information on the per cell or per user basis. On the contrary, [29] studied the power control under the assumption that the BS association is fixed. Ref. [32] studied energy saving at the BSs through striking a tradeoff between the BS association and BS operation (i.e., power control at BSs). Due to the variety of BSs in SCNs, these BS association and power control schemes proposed for macrocell-only networks cannot apply to SCNs.

Since a BS with heavy user demands is absolutely not spectrum or energy efficient due to severe interference for SCNs, it is of importance to balance traffic load and mitigate interference among BSs [36], [37]. In this direction, extensive research efforts have been put on the joint optimization of BS association and power control recently [18]–[28]. Most of previous work tend to focus on the single-objective optimization, which can be divided into two threads. The first thread is mainly concerned with minimizing the total transmit power under a predefined set of minimum SINR constraints at the receiver sides [18]–[24]. The power minimization criterion may be appropriate for networks subject to fixed rate and fixed quality-of-service (QoS) requirements. However, it fails to satisfy various concerns on the overall network performance in modern wireless networks. The second thread is concerned with maximizing the objective of the overall throughput, or more generally, a network utility function across all users in the network for modern wireless networks [25]–[28]. Although the joint BS association and power control has been studied for certain targets, the multiple access techniques used in these work are either the orthogonal multiple access (e.g., orthogonal frequency division multiple access) or the traditional non-orthogonal multiple access (i.e., pure power control).

Because of its superior spectral efficiency, NOMA with SIC has attracted lots of research interests in the literature [38]–[42]. [38] derived the capacity region for the two-user NOMA network with SIC. The evaluation of the outage probability and ergodic sum rates was carried out in [39] for a NOMA-SIC downlink cellular network with fixed power allocation.

1536-1276 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

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/TWC.2017.2664832, IEEE Transactions on Wireless Communications

Apart from the performance investigation, more research attentions have been paid to NOMA resource allocation [40]–[42]. For example, [40] maximized the sum-rate utility via the joint optimization of power and channel allocation for the downlink NOMA cellular network with SIC. [41] maximized the minimum achievable user rate through appropriate power allocation among users for the downlink NOMA cellular network with SIC. In the uplink NOMA cellular network with SIC, the optimal power allocation was studied for the maximum throughput scheduling and proportional fairness scheduling in [42]. This drives us to propose a novel non-orthogonal multiple access scheme (i.e., NOMA with SIC) to further improve the network performance of SCNs.

To strike a balance between the system-wide utility maximization and power minimization, we aim to maximize the system-wide utility and meanwhile minimize the total power consumption through the joint optimization of BS association and power control for the uplink SCN using NOMA with SIC. Due to the non-convex combinatorial nature, the optimal solution to the joint BS association and power control is difficult to obtain in the single-in single-out SCN with perfect CSI even for the single-objective optimization. Therefore, most of previous work proposed heuristic algorithms to optimize the BS association and transmit power levels. For example, [43], [44] solved the joint optimization of BS association and power control by the greedy method. [45] randomly assigned MUs to the right BSs with the probability proportional to the estimated throughput. [27] devised the algorithms based on the relaxation heuristic. [28] addressed the joint optimization problem based on the duality theory. In addition to designing heuristic algorithms, [25], [26] addressed the joint optimization of BS association and power control from a game theory perspective. However, from a system-wide utility maximization or load-balancing perspective, these games can only converge to a Nash equilibrium point, which has no guarantee to be unique or optimal and is far from being adequate. Therefore, we aim at finding a globally optimal solution to the joint BS association and power control problem for the target of maximizing the system-wide utility and meanwhile minimizing the total transmit power. Only by doing so can we use the optimal solution as a benchmark to evaluate the performance of existing heuristics for the same target, and further understand what an optimal joint BS association and power control scheme should be, how far the proposed schemes operate from the optimal performance, and how to improve the practical heuristics according to the characteristics of the optimal solution.

III. System Model and Problem Formulation

We consider an uplink SCN consisting of a set $\mathcal{N} = \{1, 2, \cdots, N_M, N_M + 1, \cdots, N\}$ of BSs and a set $\mathcal{M} = \{1, 2, \cdots, M\}$ of MUs, where N_M and $N - N_M$ mean the number of macro-cell BSs (MBSs) and small-cell BSs (SBSs), respectively. Both the BSs and MUs are assumed to be equipped with a single antenna each. We consider that the whole band of frequency can be reused by all MUs with the frequency reuse factor being one. Let the total bandwidth allocated to the uplink SCN be B. Let p_i denote the transmit power of MU i when it communicates with some BS, with $P_{i,\max}$ being its maximum allowable value. Suppose that each BS j has the perfect knowledge of all channel information through tracking the pilot signals from MUs. Correspondingly, we use the channel gain g_{ij} to denote the channel state information between MU i and BS j, which is generally determined by various factors such as path loss, shadowing and fading effects. To simplify the problem, we assume that the channel gains are frequency-flat across the whole bandwidth, and remain unchanged during the whole transmission period. We denote the set of MUs associated with BS j as $\mathcal{S}_j$. Since only MUs in the coverage of BS j are allowed to be associated with BS j in SCNs, the set $\mathcal{S}_j$ only includes intracell MUs of BS j. Suppose that the transmissions operate in the symbol time, and each BS can synchronize the transmissions of its associated MUs in each symbol time. In this case, each BS is equipped with a SIC receiver for performing the NOMA scheme with SIC. Specifically, the SIC receiver on BS j decodes the symbol signals of MUs associated with BS j in each symbol time while treating the signals of MUs associated with other BSs as interference. In each symbol time, BS j randomly selects one permutation of cancellation order set $\{1, \cdots, |\mathcal{S}_j|\}$, say $\boldsymbol{\pi}_j$, to decode the received symbols sequentially when it receives the symbols from its associated MUs, where $|\mathcal{S}_j|$ means the cardinality of $\mathcal{S}_j$. Therefore, the instantaneous data rate of MU i associated with BS j can be expressed as

$$r_{i\boldsymbol{\pi}_j} = B\log\left(1 + \frac{p_i g_{ij}}{I_{\text{intra-cell}} + I_{\text{inter-cell}} + n_j}\right), \tag{1}$$

where n_j denotes the thermal noise power at BS j, $I_{\text{intra-cell}}$ and $I_{\text{inter-cell}}$ represent the out-of-cell interference and in-cell interference of MU i. Since BS j performs the SIC to decode the signals of its associated MUs, the in-cell interference of MU i is calculated as the sum signal power from MUs associated with BS j but decoded after MU i, i.e.,

$$I_{\text{intra-cell}} = \sum_{k\in\mathcal{S}_j:\pi_{kj}>\pi_{ij}} p_k g_{kj}, \tag{2}$$

where π_{ij} is denoted as the cancellation order of MU i in the SIC permutation $\boldsymbol{\pi}_j$. Since BS j treats the signals of MUs not associated with BS j as interference, the out-of-cell interference $I_{\text{inter-cell}}$ can be expressed as

$$I_{\text{inter-cell}} = \sum_{j'=1,j'\neq j}^{N} \sum_{i'\in\mathcal{S}_{j'}} p_{i'} g_{i'j}. \tag{3}$$

We use $q_{\boldsymbol{\pi}_j}$ to denote the probability of choosing the ordering permutation $\boldsymbol{\pi}_j$ during the whole transmission period. We use $\boldsymbol{\Pi}_j$ to denote all permutations of cancellation order set $\{1, \cdots, |\mathcal{S}_j|\}$. We use x_{ij} to denote the binary variable that indicates whether MU i is associated with BS j for its load traffic. In particular, x_{ij} is allowed to be one only when MU i is associated with BS j at a satisfactory QoS level (e.g., the minimum average-rate requirement in this paper), and zero otherwise. Correspondingly, the average rate of MU i can be expressed as

$$r_i = \sum_{j=1}^{N} x_{ij} \Big(\sum_{\forall\boldsymbol{\pi}_j\in\boldsymbol{\Pi}_j} r_{i\boldsymbol{\pi}_j} q_{\boldsymbol{\pi}_j} \Big), \tag{4}$$

1536-1276 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

and the satisfactory QoS level of MU i can be evaluated as

$$r_i \geq r_{i,\min}(\sum_{j=1}^{N} x_{ij}), \quad (5)$$

where $r_{i,\min}$ represents the minimum average-rate requirement of MU i when it is served. Without loss of generality, we assume that $q_{\boldsymbol{\pi}_j} = \frac{1}{|\mathcal{S}_j|!}$ for all $\boldsymbol{\pi}_j$ in the following sections, where $(\cdot)!$ means the number of possible permutations of $|\mathcal{S}_j|$ objects from the set $\mathcal{S}_j$.

Suppose that when MU i is associated with SBS j, SBS j obtains a revenue of λ_i, where the revenue λ_i means the payment of MU i for the service provided by SBS j. For example, the payment of MU i can be interpreted as the service fee set by the network operator. The overall revenue of BS j is calculated as the weighted sum of the number of MUs associated with BS j, i.e., $\sum_{i=1}^{M} \lambda_i x_{ij}$. In the special case where λ_i is equal to 1 for all MU i, $\sum_{i=1}^{M} \lambda_i x_{ij}$ is the total number of MUs associated with BS j. For the purpose of load balancing and interference management, all MUs have to flexibly schedule their traffic among BSs. In this direction, we aim at *maximizing the system-wide utility and meanwhile minimizing the total transmit power subject to the maximum utility while meeting the minimum average-rate requirement of each served MU* through jointly optimizing the BS association and power allocation of each MU in the network. In particular, the system utility is considered as the sum of individual utilities across BSs, where the individual utility of BS j is formulated as an increasing concave function of total revenue obtained by BS j, i.e., $U_j(\sum_{i=1}^{M} \lambda_i x_{ij})$. For the brevity, we write $\boldsymbol{X} = [x_{ij}]$ and $\boldsymbol{p} = [p_i]$ as the BS association matrix and the transmit power vector, respectively. Therefore, the joint BS association and power control problem for uplink SCNs is formulated as

$$\textbf{P1-1:} \quad U^* = \max_{(\boldsymbol{p},\boldsymbol{X})} \sum_{j=1}^{N} U_j(\sum_{i=1}^{M} \lambda_i x_{ij}) \quad (6a)$$

$$s.t. \quad r_i \geq r_{i,\min}(\sum_{j=1}^{N} x_{ij}), \quad (6b)$$

$$\sum_{j=1}^{N} x_{ij} \leq 1, \quad (6c)$$

$$\begin{cases} x_{ij} \in \{0,1\}, \text{ if } i \in \mathcal{C}_j, \\ x_{ij} = 0, \text{ otherwise}, \end{cases} \quad (6d)$$

$$0 \leq p_i \leq P_{i,\max}, \forall i \in \mathcal{M}, \forall j \in \mathcal{N}, \quad (6e)$$

$$\textbf{P1-2:} \quad \min_{(\boldsymbol{p},\boldsymbol{X})} \sum_{i=1}^{M} p_i \quad (6f)$$

$$s.t. \quad \sum_{j=1}^{N} U_j(\sum_{i=1}^{M} \lambda_i x_{ij}) = U^*, \quad (6g)$$

$$\text{(6b)-(6e)}, \quad (6h)$$

where $\mathcal{C}_j$ denotes the set of MUs in the coverage of BS j. Specifically, the constraint (6b) means that to guarantee the service quality, the obtained average rate has to satisfy the minimum average-rate requirement for MU i in service. The constraint (6c) ensures that every MU can be served by at most one BS. The constraint (6e) means that MU i can be associated with BS j only when it is within the coverage of BS j. The second-stage subproblem (6f) means that every served MU should be associated with the right BS with the minimum total transmit power when the system utility is maximized. In the practical SCN, different types of base stations have different coverage, and the coverage of MBSs is generally much larger than that of SBSs.

Noticeably, we can strike different balances between the overall revenue and fairness among SBSs by appropriately choosing the utility function $U_j(\sum_{i=1}^{M} \lambda_i x_{ij})$. For example, one class of utility functions correspond to

$$U_j(\sum_{i=1}^{M} \lambda_i x_{ij}) = \begin{cases} \frac{c_j}{1-\alpha}(\sum_{i=1}^{M} \lambda_i x_{ij})^{1-\alpha}, & \alpha \geq 0 \text{ and } \alpha \neq 0 \\ c_j \log(\sum_{i=1}^{M} \lambda_i x_{ij}), & \alpha = 1, \end{cases} \quad (7)$$

where α is a fairness parameter and c_j is the QoS weight of SBS j [46], [47]. Specifically, assuming that the QoS weights are equal for all SBSs, $\alpha = 0$, $\alpha = 1$ and $\alpha \to \infty$ result in the maximum of sum-revenue, the proportionally fairness, and the max-min fairness, respectively. This implies that increasing α leads to fairer load distribution among SBSs.

We note that Problem **P1** is a mixed integer non-convex optimization problem due to the products of optimization variables (i.e., either $p_i x_{ij}$ or $p_i x_{ij'}$) in the constraints. Furthermore, it can be seen that there is an additional optimization problem in the constraint for Problem **P1**. Thus, it is difficult to solve Problem **P1** based on the convex optimization arguments. To efficiently solve it, it is significant to combine the two inter-related optimization problems into a single-stage optimization problem via weighted summing the system-wide utility and the total transmit power consumption.

Based on the idea of single-objective formulation technique in [48], we transform Problem **P1** into a single-stage optimization problem as follows:

$$\textbf{P2:} \quad \min_{(\boldsymbol{p},\boldsymbol{X})} \theta \sum_{i=1}^{M} p_i - (1-\theta) \sum_{j=1}^{N} U_j(\sum_{i=1}^{M} \lambda_i x_{ij}) \quad (8a)$$

$$s.t. \quad \sum_{\boldsymbol{\pi}_j \in \boldsymbol{\Pi}_j} q_{\boldsymbol{\pi}_j} r_{i\boldsymbol{\pi}_j} x_{ij} \geq r_{i,\min} x_{ij} \quad (8b)$$

$$\sum_{j=1}^{N} x_{ij} \leq 1, x_{ij} \in \{0,1\}, \quad (8c)$$

$$0 \leq p_{ij} \leq P_{i,\max}, \forall i \in \mathcal{M}, \forall j \in \mathcal{N}, \quad (8d)$$

where θ is a constant satisfying

$$0 \leq \theta \leq \frac{\min_{i,j}\{U_j'(\sum_{i=1}^{M} \lambda_i)\lambda_i\}}{\sum_{i=1}^{M} P_{i,\max} + \min_{i,j}\{U_j'(\sum_{i=1}^{M} \lambda_i)\lambda_i\}}. \quad (9)$$

Here, $U_j'(\sum_{i=1}^{M} \lambda_i)$ is the first-order derivative of $U_j(r_j)$ over r_j with $r_j = \sum_{i=1}^{M} \lambda_i$. Furthermore, the following proposition

1536-1276 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

shows that given θ as in (9), we can equal Problem **P1** to Problem **P2**.

Proposition 1. Given θ as in (9), the optimal solution to Problem **P2** optimizes Problem **P1**.

Proof: Let $(\boldsymbol{p}^*, \boldsymbol{X}^*)$ be the optimal solution to Problem **P1**. Let $(\boldsymbol{p}', \boldsymbol{X}')$ be any feasible solution of Problem **P2**. Since the constraints (4c)-(4e) of Problem **P1** is the same as the constraints (6b)-(6d) of Problem **P2**, $(\boldsymbol{p}^*, \boldsymbol{X}^*)$ is also the feasible solution of Problem **P2**. For the notational convenience in the following proof, we let $\Delta U = \sum_{j=1}^{N} U_j\big(\sum_{i=1}^{M} \lambda_i x_{ij}^*\big) - \sum_{j=1}^{N} U_j\big(\sum_{i=1}^{M} \lambda_i x_{ij}'\big)$ and $\Delta P = \sum_{i=1}^{M} p_i^* - \sum_{i=1}^{M} p_i'$. Due to the increasing concavity of $U_j(\cdot)$, we have

$$\Delta U \geq \min_{i,j}\big\{U_j'(\sum_{i=1}^{M} \lambda_i)\lambda_i\big\}. \tag{10}$$

This implies that

$$\begin{aligned}
&\theta\Delta P - (1-\theta)\Delta U \\
&\leq \frac{\min_{i,j}\big\{U_j'(\sum_{i=1}^{M}\lambda_i)\lambda_i\big\}\Delta P - \sum_{i=1}^{M} P_{i,\max}\Delta U}{\sum_{i=1}^{M} P_{i,\max} + \min_{i,j}\big\{U_j'(\sum_{i=1}^{M}\lambda_i)\lambda_i\big\}} \\
&\leq \frac{\min_{i,j}\big\{U_j'(\sum_{i=1}^{M}\lambda_i)\lambda_i\big\}(\Delta P - \sum_{i=1}^{M} P_{i,\max})}{\sum_{i=1}^{M} P_{i,\max} + \min_{i,j}\big\{U_j'(\sum_{i=1}^{M}\lambda_i)\lambda_i\big\}} = \Delta.
\end{aligned} \tag{11}$$

Since the value of ΔP is in $[-\sum_{i=1}^{M} P_{i,\max}, \sum_{i=1}^{M} P_{i,\max}]$, we further get from (11) that

$$\Delta \leq 0. \tag{12}$$

Together with the fact that $(\boldsymbol{p}^*, \boldsymbol{X}^*)$ is the feasible solution of Problem **P2**, it follows that $(\boldsymbol{p}^*, \boldsymbol{X}^*)$ is also the optimal solution to Problem **P2**. Hence, Proposition 1 follows immediately. ■

Proposition 1 implies that we can achieve the same performance as expressed Problem **P1** through solving Problem **P2**, as long as we tune the weight θ in the range given by (9). Specially, we note that Problem **P2** reduces to the pure system utility maximization problem when $\theta = 0$, in which there is no need to minimize the total power consumption.

Notably, Problem **P2** is now a mixed-integer non-convex single-objective optimization problem. Due to the existence of integer variables (i.e., x_{ij}'s) and continuous variables (i.e., p_i's), Problem **P2** is essentially combinatorial optimizations related to the joint BS association and power control. Thus, we show in the next section that Problem **P2** can be solved efficiently by the theory of coalition formation games and the idea of simulated annealing.

IV. The PCSUM Algorithm

In this section, we propose a novel Power Controlled System-wide Utility Maximization algorithm, referred to as PCSUM, to efficiently solve Problem **P2** based on its nature. The key idea of the PCSUM algorithm largely comes from the theory of coalition formation game [49]. For the convenience of readers, we first review some preliminaries on coalition formation games in Subsection III-A before presenting the PCSUM algorithm in Subsection III-D.

A. Preliminaries Related to Coalition Formation Games

A coalitional formation game is formulated as a triplet $(\mathcal{S}, \nu, \mathcal{B})$ in [49]. In particular, $\mathcal{S}$ denotes a set of players who seek to form cooperative groups, i.e., coalitions, in order to strengthen their positions in the game. $\mathcal{B}$ is a partition of $\mathcal{S}$, i.e., a collection of coalitions $\mathcal{B} = \{\mathcal{S}_1, \cdots, \mathcal{S}_l\}$, such that $\mathcal{S}_j \cap \mathcal{S}_{j'} = \varnothing$ for all $j \neq j'$, and $\bigcup_{j=1}^{l} \mathcal{S}_j = \mathcal{S}$. ν is the coalition value quantifying the gain of a coalition in the game, which is defined as $\nu(\mathcal{S}_j, \mathcal{B})$ for the coalition $\mathcal{S}_j \in \mathcal{S}$. It is worth noting that the coalition value of $\mathcal{S}_j$, i.e., $\nu(\mathcal{S}_j, \mathcal{B})$, also depends on the partition of $\mathcal{B}$.

The main theme of coalitional formation games is to present a partition of the players such that the sum of coalition values of all coalitions is maximized. For this purpose, an initial partition, say $\mathcal{B}_0$, is first defined in the coalitional formation game. Then, assume a possible partition, say $\mathcal{B}_1$. If $\sum_{\forall \mathcal{S}_j \in \mathcal{B}_1} \nu(\mathcal{S}_j, \mathcal{B}_1) \leq \sum_{\forall \mathcal{S}_j \in \mathcal{B}_0} \nu(\mathcal{S}_j, \mathcal{B}_0)$, then the partition $\mathcal{B}_1$ is considered as a better partition for the players. By doing so, we find the best partition in the coalitional formation game at last. Therefore, the most important objective of coalitional formation games is to provide low-complexity algorithms for building a partition of MUs related to BS association with power control in our work.

B. Formulation as a Coalition Formation Game

Note that the objective of Problem **P2** is to obtain the optimal partition of MUs that maximizes the power controlled system-wide utility. In particular, every served MU is associated with the right SBS with the minimum total transmit power in the optimal partition of MUs. Therefore, we can formulate Problem **P2** as a coalition formation game, in which the set of players corresponds to the set of MUs, i.e., $\mathcal{M}$. Assume that the partition of MUs has the form of $\mathcal{B} = \{\mathcal{S}_1, \cdots, \mathcal{S}_N, \mathcal{S}_{N+1}\}$, where the coalition $\mathcal{S}_j$ is the set of MUs served by SBS j for all $j \leq N$, and the coalition $\mathcal{S}_{N+1}$ is the set of MUs not associated with any BS yet. Since the constraint (8c) shows that every MU is served by at most one SBS, every coalition $\mathcal{S}_j$ must satisfy $\mathcal{S}_j \cap \mathcal{S}_{j'} = \varnothing$ for all $j \neq j'$, and $\bigcup_{j=1}^{N+1} \mathcal{S}_j = \mathcal{M}$. This implies that when all MUs are served by SBSs, the coalition $\mathcal{S}_{N+1}$ is empty. Based on the objective function of Problem **P2**, we define the partition function in the following form:

$$\nu(\mathcal{S}_j, \mathcal{B}) = \begin{cases} \theta \sum_{\forall i \in \mathcal{S}_j} \hat{p}_i - (1-\theta) U_j(\sum_{\forall i \in \mathcal{S}_j} \lambda_i), \text{if } j \in \mathcal{N}, \\ 0, \text{otherwise, i.e., } j = N+1, \end{cases} \tag{13}$$

where $\hat{p}_i$ means the optimal transmit power of MU i under the partition $\mathcal{B}$.

It is clear that given the partition of MUs, i.e., $\mathcal{B}$, the SBS association is then determined for every MU. In particular, if

1536-1276 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

MU i belongs to the coalition $\mathcal{S}_j$ ($\forall j \in \mathcal{N}$), then $x_{ij} = 1$, and $x_{ij'} = 0$ for all $j \neq j'$ otherwise. However, the optimal transmit power from MU i to SBS j (i.e., $\hat{p}_i$) is still unknown, even the partition of MUs is given. This implies that before calculating the partition function $\nu(\mathcal{S}_j, \mathcal{B})$, we first need to obtain the optimal transmit power $\hat{p}_i$'s for Problem **P2** based on the partition $\mathcal{B}$. In the following, we will introduce how to obtain $\hat{p}_i$'s from a given partition $\mathcal{B}$.

C. Achievement of Transmit Power

Given the partition $\mathcal{B} = \{\mathcal{S}_1, \cdots, \mathcal{S}_N, \mathcal{S}_{N+1}\}$, we have the BS association matrix $\boldsymbol{X} = [x_{ij}]_{i\in\mathcal{M}, j\in\mathcal{N}}$ satisfying

$$x_{ij} = \begin{cases} 1, & \text{if } i \in \mathcal{S}_j, \\ 0, & \text{otherwise}. \end{cases} \tag{14}$$

Once the BS association matrix $\boldsymbol{X}$ is determined, the optimal power allocation of Problem **P2** can be obtained as follows. If MU i is not associated with any SBS (i.e., $i \in \mathcal{S}_{N+1}$), then its transmit power p_i is equal to 0. Otherwise, we can obtain other $\hat{p}_i$'s (i.e., all $i \notin \mathcal{S}_{N+1}$) through solving the following total transmit power minimization problem:

$$\textbf{P3:} \quad \min_{p_i's} \sum_{\forall i \notin \mathcal{S}_{N+1}} p_i \tag{15a}$$

$$s.t. \quad \sum_{\boldsymbol{\pi}_j \in \boldsymbol{\Pi}_j : i \in \mathcal{S}_j} q_{\boldsymbol{\pi}_j} r_{i\boldsymbol{\pi}_j} \geq r_{i,\min}, \tag{15b}$$

$$0 \leq p_i \leq P_{i,\max}, \forall i \in \mathcal{S}_j, \forall j \in \mathcal{N}. \tag{15c}$$

The following Theorem 1 shows the convex nature of Problem **P3**.

Theorem 1. Problem **P3** can be turned into a convex optimization problem through the one-to-one logarithmic domain and range transformation.

Proof Sketch: Consider $r_{i\boldsymbol{\pi}_j}$'s as new variables, and rewrite Problem **P3** as

$$\textbf{P4:} \quad \min \sum_{\forall i \notin \mathcal{S}_{N+1}} p_i \tag{16a}$$

$$s.t. \quad \sum_{\boldsymbol{\pi}_j \in \boldsymbol{\Pi}_j : i \in \mathcal{S}_j} q_{\boldsymbol{\pi}_j} r_{i\boldsymbol{\pi}_j} \geq r_{i,\min}, \tag{16b}$$

$$B \log\left(1 + \frac{p_i g_{ij}}{I_{\text{intra-cell}} + I_{\text{inter-cell}} + n_j}\right) \geq r_{i\boldsymbol{\pi}_j}, \tag{16c}$$

$$\text{vars.:} \quad 0 \leq p_i \leq P_{i,\max}, r_{i\boldsymbol{\pi}_j} \geq 0, \forall i \in \mathcal{S}_j, \forall j \in \mathcal{N}. \tag{16d}$$

Note that the objective and constraint functions in Problem **P3** are all monotonically increasing functions of p_i's. It holds that the optimal solution to Problem **P4** will always occur when these inequalities are satisfied with equality. This implies that the constraint (16c) in Problem **P4** is not a relaxation, and Problem **P4** is equivalent to Problem **P3**. We then prove the convex nature of Problem **P4** by the logarithmic domain and range transformation in Appendix A.

We note that Problem **P4** has to be solved over the SIC ordering, and thus it is important to reduce the number of variables $r_{i\boldsymbol{\pi}_j}$'s for the efficient solution by classifying ordering permutations. From the constraint (16c), given any two ordering permutations $\boldsymbol{\pi}_j$ and $\bar{\boldsymbol{\pi}}_j$, we have $r_{i\boldsymbol{\pi}_j} = r_{i\bar{\boldsymbol{\pi}}_j}$ according to the SINR term in (16c) if $\{k | k \in \mathcal{S}_j : \pi_{kj} > \pi_{ij}\}$ coincides with $\{k | k \in \mathcal{S}_j : \bar{\pi}_{kj} > \bar{\pi}_{ij}\}$. This implies that we can reduce the number of variables $r_{i\boldsymbol{\pi}_j}$'s by classifying all ordering permutations which can achieve the same data rate for MU i into a set. For the brevity, we re-index the MU i served by BS j as j_i, and we sort all MUs served by BS j in the increasing order of index i. We then use an $(|S_j| - 1)$-bit binary number (say $\boldsymbol{b}_{j_i}$) in the form of $\boldsymbol{b}_{j_i} = (b_{j_i j_{i'}})_{\forall i' \in \mathcal{S}_j \setminus i}$ to represent the successive cancellation orders of the other MUs associated with BS j. Specifically, each bit $b_{j_i j_{i'}}$ is equal to zero if MU $j_{i'}$ is decoded after MU j_i, and it is equal to one otherwise. Therefore, the set of ordering permutations that achieve the same rate for MU j_i can be expressed as

$$\boldsymbol{\Pi}_{\boldsymbol{b}_{j_i}} = \{\boldsymbol{\pi}_j | \pi_{j_{i'} j} > \pi_{j_i j} : b_{j_i j_{i'}} = 0, \pi_{j_{i'} j} < \pi_{j_i j} : b_{j_i j_{i'}} = 1, \text{ and } \boldsymbol{\pi}_j \in \boldsymbol{\Pi}_j\}. \tag{17}$$

Correspondingly, the probability of choosing $\boldsymbol{\Pi}_{\boldsymbol{b}_{j_i}}$ can be expressed as

$$q_{\boldsymbol{b}_{j_i}} = \sum_{\boldsymbol{\pi}_j \in \boldsymbol{\Pi}_{\boldsymbol{b}_{j_i}}} q_{\boldsymbol{\pi}_j} = \frac{N_{\text{zero}}!(|\mathcal{S}_j| - 1 - N_{\text{zero}})!}{|\mathcal{S}_j|!}, \tag{18}$$

where N_{zero} represents the number of zeros in the $(|\mathcal{S}_j| - 1)$-bit binary number $\boldsymbol{b}_{j_k}$. We use $\hat{r}_{\boldsymbol{b}_{j_i}}$ to denote the data rate of MU j_i under the ordering permutations belonging to $\boldsymbol{\Pi}_{\boldsymbol{b}_{j_i}}$. Therefore, we can rewritten Problem **P4** as

$$\textbf{P5:} \quad \min_{p_{j_i}'s} \sum_{j=1}^{N} \sum_{\forall i \in \mathcal{S}_j} p_{j_i} \tag{19a}$$

$$s.t. \quad \sum_{\boldsymbol{b}_{j_i}=0}^{2^{|\mathcal{S}_j|-1}-1} \hat{r}_{\boldsymbol{b}_{j_i}} q_{\boldsymbol{b}_{j_i}} \geq r_{j_i,\min}/B, \tag{19b}$$

$$(\boldsymbol{I} - \boldsymbol{Q})\boldsymbol{p} \geq \boldsymbol{u}, \tag{19c}$$

$$\text{vars.:} \quad 0 \leq p_{j_i} \leq P_{j_i,\max}, r_{\boldsymbol{b}_{j_i}} \geq 0, \forall i \in \mathcal{S}_j, \forall j \in \mathcal{N}. \tag{19d}$$

Here, the matrix $\boldsymbol{I}$ is the $\sum_{j=1}^{N} |\mathcal{S}_j| 2^{|\mathcal{S}_j|-1} \times \sum_{j=1}^{N} |\mathcal{S}_j|$ matrix with every entry being

$$I_{(l+\boldsymbol{b}_{j_i}+1)n} = \begin{cases} 1, & \text{if } n = \sum_{m=1}^{j-1} |\mathcal{S}_m| + |\mathcal{I}_{j_i}|, \\ 0, & \text{otherwise}, \end{cases} \tag{20}$$

where $\mathcal{I}_{j_i} = \{i' | i' \leq i \text{ and } i' \in \mathcal{S}_j\}$, and $l = \sum_{m=1}^{j-1} |\mathcal{S}_m| 2^{|\mathcal{S}_m|-1} + |\mathcal{I}_{j_i}| 2^{|\mathcal{S}_m|-1}$. The matrix $\boldsymbol{Q}$ is a

1536-1276 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

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/TWC.2017.2664832, IEEE Transactions on Wireless Communications

$\sum_{j=1}^{N} |\mathcal{S}_j| 2^{|\mathcal{S}_j|-1} \times \sum_{j=1}^{N} |\mathcal{S}_j|$ matrix satisfying

$$Q_{(l+\boldsymbol{b}_{j_i}+1)n} = \begin{cases} \frac{(2^{\hat{r}_{\boldsymbol{b}_{j_i}}}-1)g_{j'_n j}}{g_{j_i j}}, \text{if } n \leq \sum_{m=1}^{j-1} |\mathcal{S}_m| \text{ or } n \geq \sum_{m=1}^{j} |\mathcal{S}_m| + 1, \\ \frac{(2^{\hat{r}_{\boldsymbol{b}_{j_i}}}-1)g_{j_{i'} j}}{g_{j_i j}}, \text{if } n = \sum_{m=1}^{j-1} |\mathcal{S}_m| + |\mathcal{I}_{j_{i'}}| \text{ and } b_{j_i j_{i'}} = 0, \\ 0, \text{otherwise}, \end{cases} \tag{21}$$

The vector $\boldsymbol{u}$ is a $\sum_{j=1}^{N} |\mathcal{S}_j| 2^{|\mathcal{S}_j|-1} \times 1$ vector with every element being

$$u_{l+\boldsymbol{b}_{j_i}+1} = \frac{(2^{\hat{r}_{\boldsymbol{b}_{j_i}}}-1)n_j}{g_{j_i j}}. \tag{22}$$

After reducing the number of variables, we want to solve Problem **P5** (and hence Problem **P4**) in the following. Due to the convex nature, the feasibility checking can be performed via efficient interior-point algorithms proposed in [52]. If it is feasible, the primal decomposition can be used to solve Problem **P5** by separating its optimization into two levels of optimization [53] and then alternatively optimizing variables $\hat{r}_{\boldsymbol{b}_{j_i}}$'s and p_{j_i}'s in the two levels of optimization. At the lower level, we have the subproblem that optimizes variables p_{j_i}'s by solving Problem **P5** when $r_{\boldsymbol{b}_{j_i}}$'s are fixed, i.e.,

$$\begin{aligned} \min_{p_{j_i}'s} & \sum_{j=1}^{N} \sum_{\forall i \in \mathcal{S}_j} p_{j_i} \\ s.t. \quad & (\boldsymbol{I}-\boldsymbol{Q})\boldsymbol{p} \geq \boldsymbol{u}, \\ & 0 \leq p_{j_i} \leq P_{j_i,\max}, \forall j_i \notin \mathcal{S}_{N+1}. \end{aligned} \tag{23}$$

At the higher level, we have the master problem that optimizes the coupling variables $\hat{r}_{\boldsymbol{b}_{j_i}}$'s with the updated p_{j_i}'s, i.e.,

$$\begin{aligned} \min_{\hat{\boldsymbol{r}}} & \sum_{j=1}^{N} \sum_{\forall i \in \mathcal{S}_j} \hat{p}_{j_i}(\hat{\boldsymbol{r}}) \\ s.t. \quad & \sum_{\boldsymbol{b}_{j_i}=0}^{2^{|\mathcal{S}_j|-1}-1} \hat{r}_{\boldsymbol{b}_{j_i}} q_{\boldsymbol{b}_{j_i}} \geq r_{j_i,\min}/B, \forall j_i \in \mathcal{S}_j, \forall j \in \mathcal{N}, \end{aligned} \tag{24}$$

where $\hat{\boldsymbol{r}}$ are the vector of $\hat{r}_{\boldsymbol{b}_{j_i}}$'s, and $\hat{p}_{j_i}(\hat{\boldsymbol{r}})$ is the objective value of problem (23) for given $\hat{\boldsymbol{r}}$.

Due to the convex nature of Problem **P5**, problems (23) and (24) are both convex programs. In the following, we obtain the optimal solution to Problem **P5** by alternately solving problems (23) and (24). We first focus on the solution to problem (23). Since problem (23) is a linear programming, we can solve it via existing simplex methods or interior-point algorithms [52]. Assume the optimal power allocation and rate allocation to be $\hat{\boldsymbol{p}}(t)$ and $\hat{\boldsymbol{r}}(t)$ at the tth iteration, respectively. Let diag$(\cdot)$ denote diagonalization, and $[\cdot]_{\mathcal{R}}$ denote the projection onto the feasible convex set $\mathcal{R} = \left\{ \hat{\boldsymbol{r}} : \sum_{\boldsymbol{b}_{j_i}=0}^{2^{|\mathcal{S}_j|-1}-1} \hat{r}_{\boldsymbol{b}_{j_i}} q_{\boldsymbol{b}_{j_i}} \geq r_{j_i,\min}/B, \forall j_i \in \mathcal{S}_j, \forall j \in \mathcal{N} \right\}$. Due to the convex nature, the master problem (24) can be solved with a projected subgradient method by updating $\hat{\boldsymbol{r}}$ as

$$\hat{\boldsymbol{r}}(t+1) = \left[\hat{\boldsymbol{r}}(t) + \alpha_t \text{diag}(\boldsymbol{\lambda}^*(t))(\hat{\boldsymbol{Q}}\hat{\boldsymbol{p}}(t) - \hat{\boldsymbol{u}}) \right]_{\mathcal{R}}, \tag{25}$$

where $[\cdot]_{\mathcal{R}}$ means the Euclidean projection on $\mathcal{R}$, $\alpha_t > 0$ is the tth step size, the matrix $\hat{\boldsymbol{Q}}$ respectively satisfies

$$\hat{Q}_{(l+\boldsymbol{b}_{j_i}+1)n} = \begin{cases} \frac{2^{\hat{r}_{\boldsymbol{b}_{j_i}}(t)} g_{j'_n j} \ln 2}{g_{j_i j}}, \text{if } n \leq \sum_{m=1}^{j-1} |\mathcal{S}_m| \text{ or } n \geq \sum_{m=1}^{j} |\mathcal{S}_m| + 1, \\ \frac{2^{\hat{r}_{\boldsymbol{b}_{j_i}}(t)} g_{j_{i'} j} \ln 2}{g_{j_i j}}, \text{if } n = \sum_{m=1}^{j-1} |\mathcal{S}_m| + |\mathcal{I}_{j_{i'}}| \text{ and } b_{j_i j_{i'}} = 0, \\ 0, \text{otherwise}, \end{cases} \tag{26}$$

the vector $\hat{\boldsymbol{u}}$ satisfies

$$\hat{u}_{l+\boldsymbol{b}_{j_i}+1} = \frac{2^{\hat{r}_{\boldsymbol{b}_{j_i}}(t)} n_j \ln 2}{g_{j_i j}}, \tag{27}$$

and $\boldsymbol{\lambda}^*(t)$ is the optimal multiplier vector corresponding to the first constraint in problem (23) at the tth iteration.

Summarizing, we can now obtain the optimal transmit power of Problem **P2** for a given partition $\mathcal{B} = \{\mathcal{S}_1, \cdots, \mathcal{S}_N, \mathcal{S}_{N+1}\}$ (i.e., Problem **P3**) via Procedure 1.

Procedure 1 Achieve the optimal transmit power of each MU for a given partition $\mathcal{B} = \{\mathcal{S}_1, \cdots, \mathcal{S}_N, \mathcal{S}_{N+1}\}$

1: Obtain the BS association matrix $\boldsymbol{X}$ as shown in (14) by the given partition $\mathcal{B}$.
2: Let $\hat{p}_i = 0$ such that $x_{ij} = 0$ for all j. For all i and j such that $x_{ij} = 1$, re-index the MU i served by SBS j as j_i, and sort all MUs served by SBS j in the increasing order of index i.
3: **Initialization**: Check the feasibility of Problem **P5** via the interior-point method. If Problem **P5** is infeasible, BSs cannot serve all MUs in the partition $\mathcal{B}$ with feasible transmit power, and we need to find a feasible partition instead. Otherwise, set $t = 1$ and $\hat{\boldsymbol{r}}(1)$ equal to some feasible solution in $\mathcal{R}$.
4: Obtain $\hat{\boldsymbol{p}}(t)$ by solving the subproblem (23) via the interior-point method when setting $\hat{\boldsymbol{r}} = \hat{\boldsymbol{r}}(t)$.
5: Update $\hat{\boldsymbol{r}}(t+1)$ by (25).
6: Set $t \leftarrow t+1$ and go to step 4 (until satisfying the current minimum, i.e., $\min_{t_0=1,\cdots,t-1} \sum_{j=1}^{N} \sum_{\forall i \in \mathcal{S}_j} \hat{p}_{j_i}(\hat{\boldsymbol{r}}(t_0))$ is less than ϵ).
7: If not all served MUs have feasible transmit power in the partition $\mathcal{B}$, and we need to find a feasible partition instead. Otherwise, we have the optimal power allocation of MUs served in the partition $\mathcal{B}$ to be $\hat{\boldsymbol{p}} = \hat{\boldsymbol{p}}(t-1)$.

D. The PCSUM Algorithm

The PCSUM algorithm works as follows.

1536-1276 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

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/TWC.2017.2664832, IEEE Transactions on Wireless Communications

1) Initialization

Recall that the key idea of the PCSUM algorithm is to find an optimal partition of MUs corresponding to the optimal solution to Problem **P1** from an feasible initial partition of MUs. Thus, we first need to choose an initial feasible partition $\mathcal{B}^{(0)} = \{\mathcal{S}_1^{(0)}, \cdots, \mathcal{S}_N^{(0)}, \mathcal{S}_{N+1}^{(0)}\}$ via the following operations. Initially, let $\mathcal{S}_j^{(0)} = \varnothing$ for all $j \in \mathcal{N}$ and $\mathcal{S}_{N+1}^{(0)} = \mathcal{M}$. Then, we finish the structure of $\mathcal{B}^{(0)}$ through M steps. In the ith step, we remove MU i from $\mathcal{S}_{N+1}^{(0)}$ to $\mathcal{S}_j^{(0)}$, which is randomly chosen in $\{\mathcal{S}_1^{(0)}, \cdots, \mathcal{S}_N^{(0)}\}$ and satisfies $i \in \mathcal{C}_j$. When MU i is included in the coalition $\mathcal{S}_j^{(0)}$, we then check whether the new $\mathcal{B}^{(0)}$ is feasible via Procedure 1. If the new $\mathcal{B}^{(0)}$ is feasible, then MU i is removed to $\mathcal{S}_j^{(0)}$ from $\mathcal{S}_{N+1}^{(0)}$. Otherwise, MU i still remains in $\mathcal{S}_{N+1}^{(0)}$. By doing so, the initial feasible partition $\mathcal{B}^{(0)}$ is obtained after M steps. In particular, we can obtain the initial partition $\mathcal{B}^{(0)}$ as shown in Procedure 2.

Procedure 2 Achieve the initial feasible partition $\mathcal{B}^{(0)} = \{\mathcal{S}_1^{(0)}, \cdots, \mathcal{S}_N^{(0)}, \mathcal{S}_{N+1}^{(0)}\}$

1: Let $\mathcal{S}_{N+1}^{(0)} = \mathcal{M}$.
2: **for** MU $i = 1 \to M$ **do**
3: Remove MU i from $\mathcal{S}_{N+1}^{(0)}$ to $\mathcal{S}_j^{(0)}$, which is randomly chosen in $\{\mathcal{S}_1^{(0)}, \cdots, \mathcal{S}_N^{(0)}\}$ and satisfies $i \in \mathcal{C}_j$.
4: Check whether the new $\mathcal{B}^{(0)}$ is feasible via the interior-point method.
5: If the new $\mathcal{B}^{(0)}$ is feasible, then MU i is removed to $\mathcal{S}_j^{(0)}$ from $\mathcal{S}_{N+1}^{(0)}$. Otherwise, MU i still remains in $\mathcal{S}_{N+1}^{(0)}$.
6: **end for**

2) Update of partition

As mentioned before, we have formulated Problem **P2** as the coalition formation game $\{\mathcal{M}, \nu, \mathcal{B}\}$. Next, based on the coalition formation game, we introduce how to iteratively update the partition of MUs in the framework of simulated annealing to obtain the optimal partition (and hence the optimal solution to Problem **P2**).

The main operations work as follows at the kth iteration. First, we randomly choose two MUs[1] between two coalitions from the partition $\mathcal{B}_{k-1}$. If the chosen MU is not within the coverage of the other BS, then we keep this MU in the current coalition and assume a virtual MU is chosen. Second, we exchange these two MUs between these two coalitions to obtain a temporal partition $\tilde{\mathcal{B}}$. Third, we calculate the BS matrix $\boldsymbol{X}$ and the optimal transmit power allocation $\hat{\boldsymbol{p}}$ corresponding to $\tilde{\mathcal{B}}$ by Procedure 1. If the partition $\tilde{\mathcal{B}}$ is infeasible, then $\tilde{\mathcal{B}}$ is rejected and $\mathcal{B}_{k-1}$ keeps as the partition $\mathcal{B}_k$ at the kth iteration. Otherwise, we calculate the total coalition value on the partition $\tilde{\mathcal{B}}$ (i.e., $\sum_{\forall \mathcal{S}_j \in \tilde{\mathcal{B}}} \nu(\mathcal{S}_j, \tilde{\mathcal{B}})$) according to the BS matrix $\boldsymbol{X}$ and the power allocation $\hat{\boldsymbol{p}}$ obtained by Procedure 1, where every term $\nu(\mathcal{S}_j, \tilde{\mathcal{B}})$ is calculated by (13). Finally, we compare the total coalition value on the partition $\tilde{\mathcal{B}}$ with that on the partition $\mathcal{B}_k$. If we have $\sum_{\forall \mathcal{S}_j \in \tilde{\mathcal{B}}} \nu(\mathcal{S}_j, \tilde{\mathcal{B}}) \le \sum_{\forall \mathcal{S}_j \in \mathcal{B}_{k-1}} \nu(\mathcal{S}_j, \mathcal{B}_{k-1})$, then $\tilde{\mathcal{B}}$ is accepted as the partition $\mathcal{B}_k$ with probability 1 at the kth iteration. Otherwise, $\tilde{\mathcal{B}}$ is accepted as the partition $\mathcal{B}_k$ with a probability of $\exp(\frac{\Delta}{T})$, and $\mathcal{B}_{k-1}$ keeps as the partition $\mathcal{B}_k$ with a probability of $1 - \exp(\frac{\Delta}{T})$ at the kth iteration, where $\Delta = \sum_{\forall \mathcal{S}_j \in \mathcal{B}_{k-1}} \nu(\mathcal{S}_j, \mathcal{B}_{k-1}) - \sum_{\forall \mathcal{S}_j \in \tilde{\mathcal{B}}} \nu(\mathcal{S}_j, \tilde{\mathcal{B}})$ and T is a control parameter (also referred to as temperature). For convergence, T decreases with each iteration. Intuitively, the acceptance of uphill-moving becomes less and less likely with the decrease of T, implying convergence when T is sufficiently small. When $\tilde{\mathcal{B}}$ is updated as $\mathcal{B}_k$, $\sum_{\forall \mathcal{S}_j \in \tilde{\mathcal{B}}} \nu(\mathcal{S}_j, \tilde{\mathcal{B}})$ keep as the memory of $\sum_{\forall \mathcal{S}_j \in \mathcal{B}_k} \nu(\mathcal{S}_j, \mathcal{B}_k)$, and the BS association matrix $\boldsymbol{X}$ and the transmit power $\hat{\boldsymbol{p}}$ corresponding to $\tilde{\mathcal{B}}$ is considered as the BS association matrix $\boldsymbol{X}_k$ and the transmit power $\hat{\boldsymbol{p}}_k$ at the kth iteration. When $\mathcal{B}_{k-1}$ keeps as $\mathcal{B}_k$, $\sum_{\forall \mathcal{S}_j \in \mathcal{B}_{k-1}} \nu(\mathcal{S}_j, \mathcal{B}_{k-1})$ keep as the memory of $\sum_{\forall \mathcal{S}_j \in \mathcal{B}_k} \nu(\mathcal{S}_j, \mathcal{B}_k)$, and the BS association matrix $\boldsymbol{X}_{k-1}$ and the transmit power $\hat{\boldsymbol{p}}_{k-1}$ obtained at the $(k-1)$th iteration keeps as the BS association matrix $\boldsymbol{X}_k$ and the transmit power $\hat{\boldsymbol{p}}_k$ at the kth iteration.

Having introduced the basic operations, we now formally present the PCSUM algorithm[2] as Algorithm 1.

E. Convergence and Complexity

Note that the implementation of the PCSUM algorithm consisting of a series of Procedure 1. Before analyzing the computational complexity, we need to analyze the computational complexity of Procedure 1. Therefore, the following proposition shows the computational complexity of Procedure 1.

Proposition 2. Let the step size be $\frac{1}{t}$ at the tth iteration. Then, the computational complexity that Procedure 1 obtains the ϵ-optimal solution[3] to Problem **P5** is of $O(\sqrt{5(A+B)}) + O(A^{0.5} e^{1/\epsilon})$ iterations. Here, $A = \sum_{j=1}^{N} |\mathcal{S}_j|$ and $B = \sum_{j=1}^{N} |\mathcal{S}_j| 2^{|\mathcal{S}_j|-1}$.

We provide details of the proof in Appendix B.

The following proposition discusses the convergence of the PCSUM algorithm.

Proposition 3 ([50], [51]). The PCSUM algorithm converges to the global optimal solutions to Problem **P2** (and hence

[1]When we choose one MU from the coalition $\mathcal{S}_j$, the chosen MU is allowed to be either a particular MU belonging to $\mathcal{S}_j$ or a virtual MU implying no MU is chosen.

[2]In SCNs, BSs are assumed to be connected to a cloud-computing server provided by the network operator via the wired backhauls. Each BS needs to collect the topology information of mobile users (e.g., the channel information between MUs and BSs), and then sends the topology information to the server. The server is responsible for running the PCSUM algorithm based on the received topology information, and then informs each BS of the joint optimal BS association and power control. Each BS further informs the served mobile user of its possible transmit power.

[3]In this paper, the ϵ-optimal solution means the feasible solution such that the minimum obtained so far is ϵ less than the optimal value.

1536-1276 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

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/TWC.2017.2664832, IEEE Transactions on Wireless Communications

Algorithm 1 The PCSUM Algorithm for solving Problem **P2** (i.e., Problem **P1**)

1: Initialization: Achieve the initial feasible partition $\mathcal{B}^{(0)} = \{\mathcal{S}_1^{(0)}, \cdots, \mathcal{S}_N^{(0)}, \mathcal{S}_{N+1}^{(0)}\}$ by Procedure 2. Let $k = 1$, and $\mathcal{B}^{(1)} = \mathcal{B}^{(0)}$.
2: **repeat**
3: Randomly choose two coalitions from the partition $\mathcal{B}_k$, say $\mathcal{S}_j$ and $\mathcal{S}_{j'}$.
4: Randomly choose two MUs, say i and i', from $\mathcal{S}_j$ and $\mathcal{S}_{j'}$, respectively.
5: Obtain a temporal partition $\tilde{\mathcal{B}}$ through exchanging two MUs i and i' between $\mathcal{S}_j$ and $\mathcal{S}_{j'}$.
6: Calculate the BS matrix $\boldsymbol{X}$ and the transmit power matrix $\hat{\boldsymbol{p}}$ corresponding to $\tilde{\mathcal{B}}$ by Procedure 1.
7: If the partition $\tilde{\mathcal{B}}$ is infeasible, then $\tilde{\mathcal{B}}$ is rejected and $\mathcal{B}_{k-1}$ keeps as the partition $\mathcal{B}_k$ at the kth iteration. Otherwise, we further calculate the total coalition value on the partition $\tilde{\mathcal{B}}$ (i.e., $\sum_{\forall \mathcal{S}_j \in \mathcal{B}_t} \nu(\mathcal{S}_j, \mathcal{B}_t)$) according to the BS matrix $\boldsymbol{X}$ and the transmit power matrix $\hat{\boldsymbol{p}}$, where every term $\nu(\mathcal{S}_j, \tilde{\mathcal{B}})$ is calculated by (13).
8: Compare $\sum_{\forall \mathcal{S}_j \in \tilde{\mathcal{B}}} \nu(\mathcal{S}_j, \tilde{\mathcal{B}})$ with $\sum_{\forall \mathcal{S}_j \in \mathcal{B}_{k-1}} \nu(\mathcal{S}_j, \mathcal{B}_{k-1})$. If we have $\sum_{\forall \mathcal{S}_j \in \tilde{\mathcal{B}}} \nu(\mathcal{S}_j, \tilde{\mathcal{B}}) \leq \sum_{\forall \mathcal{S}_j \in \mathcal{B}_{k-1}} \nu(\mathcal{S}_j, \mathcal{B}_{k-1})$, then $\tilde{\mathcal{B}}$ is accepted as the partition $\mathcal{B}_k$ with probability 1. Otherwise, $\tilde{\mathcal{B}}$ is accepted as the partition $\mathcal{B}_k$ with a probability of $\exp(\frac{\Delta}{T})$, and $\mathcal{B}_{k-1}$ keeps as the partition $\mathcal{B}_k$ with a probability of $1 - \exp(\frac{\Delta}{T})$, where $\Delta = \sum_{\forall \mathcal{S}_j \in \mathcal{B}_{k-1}} \nu(\mathcal{S}_j, \mathcal{B}_{k-1}) - \sum_{\forall \mathcal{S}_j \in \tilde{\mathcal{B}}} \nu(\mathcal{S}_j, \tilde{\mathcal{B}})$.
9: When $\tilde{\mathcal{B}}$ is updated as $\mathcal{B}_k$, $\sum_{\forall \mathcal{S}_j \in \tilde{\mathcal{B}}} \hat{\nu}(\mathcal{S}_j, \tilde{\mathcal{B}})$ and $\{\boldsymbol{X}, \hat{\boldsymbol{p}}\}$ keep as the memory of $\sum_{\forall \mathcal{S}_j \in \mathcal{B}_k} \hat{\nu}(\mathcal{S}_j, \mathcal{B}_k)$ and $\{\boldsymbol{X}_k, \hat{\boldsymbol{p}}_k\}$, respectively. When $\mathcal{B}_{k-1}$ keeps as $\mathcal{B}_k$, $\sum_{\forall \mathcal{S}_j \in \mathcal{B}_{k-1}} \hat{\nu}(\mathcal{S}_j, \mathcal{B}_{k-1})$ and $\{\boldsymbol{X}_{k-1}, \hat{\boldsymbol{p}}_{k-1}\}$ keep as the memory of $\sum_{\forall \mathcal{S}_j \in \mathcal{B}_k} \hat{\nu}(\mathcal{S}_j, \mathcal{B}_k)$ and $\{\boldsymbol{X}_k, \hat{\boldsymbol{p}}_k\}$, respectively.
10: $k = k + 1$.
11: $T = \frac{T_0}{\log(k)}$.
12: **until** $T < \eta$ (where η is a small positive number).
13: The optimal solution to Problem **P1** is equal to $(\boldsymbol{X}_{k-1}, \hat{\boldsymbol{p}}_{k-1})$.

Problem **P1**), as the control parameter T approaches to zero with $T = \frac{T_0}{\log(k)}$.

For practical implementation, a solution very close to the global optimal solution[4] is obtained when $T < \eta$.

The following proposition shows the computational complexity of the PCSUM algorithm.

Proposition 4. The computational complexity of the PCSUM algorithm is of $(\exp(\frac{T_0}{\eta})+M)(O(\sqrt{5(A+B)})+O(A^{0.5}e^{1/\epsilon}))$ iterations.

Proof: The computational complexity of the PCSUM algorithm comes from the initialization part and the repeat part. The initialization part consists of M Procedure 1's. The PCSUM algorithm converges when $\frac{T_0}{\log(k)} < \eta$, and thus $\exp(\frac{T_0}{\eta})$ Procedure 1's are needed in the repeat part. Every Procedure 1 requires a computational complexity of $O(\sqrt{5(A+B)})+O(A^{0.5}e^{1/\epsilon})$ by Proposition 2, and thus the total computational complexity has $(\exp(\frac{T_0}{\eta}) + M)(O(\sqrt{5(A+B)})+O(A^{0.5}e^{1/\epsilon}))$ iterations. ■

[4]Since Problem **P2** is a mixed-integer non-convex optimization problem, there may exist several distinct global optimal solutions. The PCSUM algorithm is proposed for finding one such solution.

TABLE I
SIMULATION PARAMETERS

Simulation parameters	**Value choice**
Carrier frequency	900 MHz
Bandwidth	20 MHz
Path loss model	$(128.1 + 37.6 \log_{10}(d))$ dB (d in km)
Noise power spectral density	-174 dBm/Hz
Maximum transmit power $P_{i,\max}$	20 dBm
Minimum average-rate requirement $r_{i,\min}$	1 Mbps

V. Simulation Results

In this section, we conduct simulations to illustrate the effectiveness of the proposed PCSUM algorithm. Following the networking parameters suggested in [54], we set the simulation parameters in Table I.

A. Optimal SBS Association under Different System-wide Utilities

Example 1: In this simulation example, we want to verify the optimal BS association under different system-wide utilities for the uplink SCN as shown in Fig. 1. Specifically, one MBS and three SBSs are respectively placed in (100m, 100m), (300m, 100m), (300m, 300m), and (100m, 300m). The coverage of MBS is an 1000-meter two-dimensional circle plane, and the coverage of each SBS is a 300-meter two-dimensional circle plane. 5 MUs, 10 MUs, 15 MUs, and 20 MUs are uniformly distributed in 100-meter two-dimensional circle planes of MBS and SBSs, respectively. We set θ to be 0.01 by (9), and and set λ_i to be 1 for all MU i.

We first consider the system-wide utility as the overall system revenue across BSs, i.e., $U_j(\sum_{i=1}^{M} \lambda_i x_{ij}) = \sum_{i=1}^{M} x_{ij}$ with $\lambda_i = 1$ for all i. In this case, the overall system revenue is evaluated as the maximal number of MUs that can be admitted into the network given the MUs' requirements, which is referred to as the equivalent capacity in [55]. Fig. 2 shows the optimal BS association that maximizes the equivalent capacity and meanwhile minimizes the total power consumption based on the maximum equivalent capacity for the network as shown in Fig. 1. In particular, it can be seen that MBS, SBS 1, SBS 2 and SBS 3 serve 5 MUs, 3 MUs, 7 MUs and 6

1536-1276 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

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/TWC.2017.2664832, IEEE Transactions on Wireless Communications

MUs, respectively. The interesting observation in Fig. 2 reveals that when maximizing the equivalent capacity with minimum total power consumption, the optimal BS association causes unbalanced load distribution among BSs. This observation results from the co-channel interference between MUs, and thus the neighboring BSs should use different spectrum for improving the load balance in uplink SCNs. From Fig. 2, we further find that not every served MU is associated with the nearest BS when maximizing the overall system revenue with minimum total power consumption. Such an observation mainly results from the following reasons. On the one hand, our work allows more than one users to be associated to each BS, and applies the NOMA with SIC to reduce intra-cell interference. Also, each user is allowed to mitigate both intra-cell and inter-cell interference with with power control in which the transmit power of each user is not proportional to the distance between the user and its associated BS any more. As a result, some users that might be far from their associated BSs would lower their transmit power due to lower intra-cell interference, which further leads to lower inter-cell interference. On the other hand, we adopt a random SIC order in each symbol time to cancel part of intra-cell interference for all users associated to the same BS, and thus we use the average rate as the performance metric of users. This implies that each user cannot achieve the same SINR target in each symbol time even they have the same minimum average-rate requirement. Therefore, under our proposed BS association strategy, some users do not connect to the nearest BSs any more by properly controlling transmit power of individual users.

We further consider the system-wide utility as the proportional fairness and the max-min fairness among SBSs with $\lambda_i = 1$ for all MU i, i.e., $U_j(\sum_{i=1}^{M} x_{ij}) = \log(\sum_{i=1}^{M} x_{ij})$ and $\sum_{j=1}^{N} U_j(\sum_{i=1}^{M} x_{ij}) = \min_j \sum_{i=1}^{M} x_{ij}$, respectively. Fig. 3 shows the optimal BS association that maximizes the proportional fairness with minimum total power consumption in the network as shown in Fig. 1, and Fig. 4 shows the optimal BS association of maximizing the max-min fairness. In the comparison of these three system utilities, we find that by the Jain's fairness index [56], the max-min fairness utility can achieve the highest load balancing among BSs, although resulting in the fewest MUs in service. This observation implies that the total number of served MUs and the fairness of load balancing among BSs are two competitive components. In other words, the fewer MUs are served if we want the fairer load balancing among BSs. Therefore, the appreciate system-wide utility function needs to be chosen if we want to strike a balance between the fairness of load balancing among BSs and the number of MUs in service.

Example 2: This simulation example is to show the effect of weight θ on the performance of the proposed algorithm. We consider the same simulation topology as Example 1. Fig. 5 shows the number of served MUs and the power consumption obtained by the proposed algorithm with different θ's for the equivalent capacity utility, the proportional fairness utility and the max-min fairness utility, respectively. It can be seen

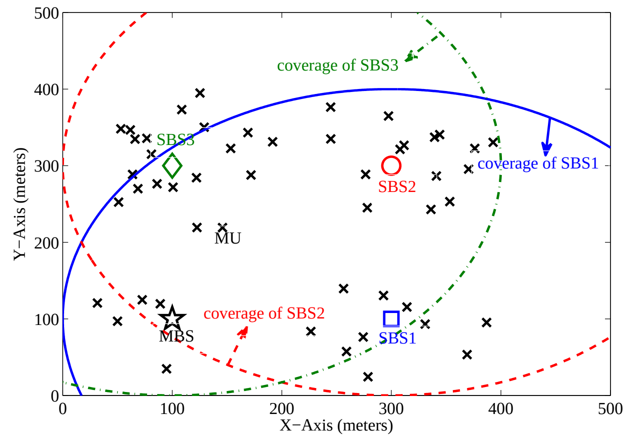


Fig. 1. An SCN with 50 MUs and 4 SBSs.

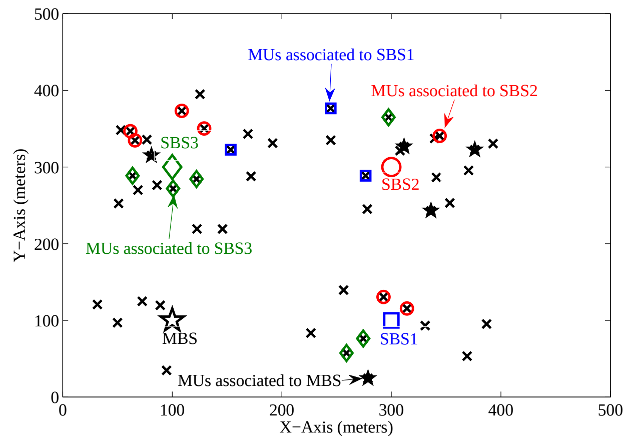


Fig. 2. The optimal SBS association obtained when maximizing the equivalent capacity (i.e., $\sum_{j=1}^{N}\sum_{i=1}^{M} x_{ij}$) for the uplink SCN as in Fig. 1. Specifically, the number of MUs in service is 5, 3, 7, and 6 associated to MBS, SBS 1, SBS 2, and SBS 3, respectively.

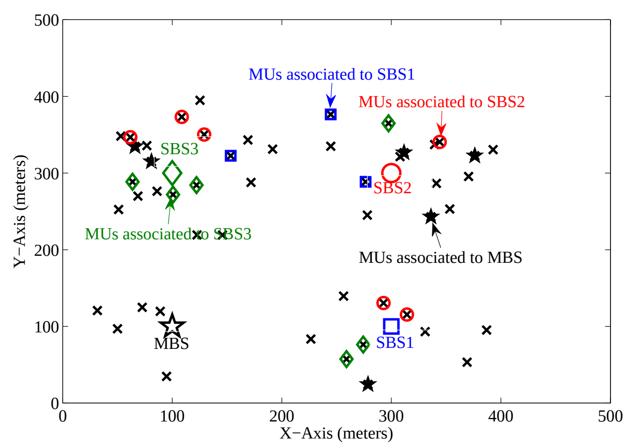


Fig. 3. The optimal SBS association obtained when maximizing the proportional fairness (i.e., $\sum_{j=1}^{N} \log(\sum_{i=1}^{M} x_{ij})$) for the uplink SCN as in Fig 1. Specifically, the number of MUs in service is 6, 3, 6, and 6 associated to MBS, SBS 1, SBS 2, and SBS 3, respectively.

1536-1276 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

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/TWC.2017.2664832, IEEE Transactions on Wireless Communications

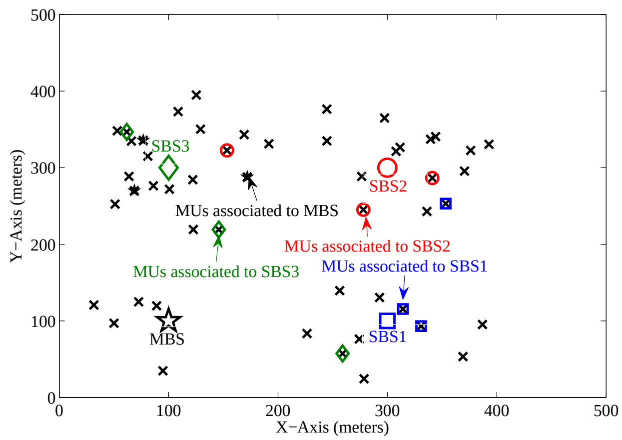


Fig. 4. The optimal SBS association obtained when maximizing the max-min fairness (i.e., $\min_{j} \sum_{i=1}^{M} x_{ij}$) for the uplink SCN as in Fig 1. Specifically, the number of MUs in service is 3, 3, 3, and 3 associated to MBS, SBS 1, SBS 2, and SBS 3, respectively.

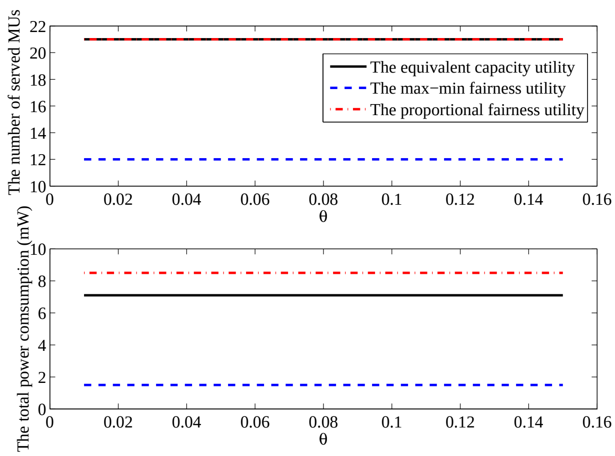


Fig. 5. The effect of weight θ on the performance of the proposed algorithm for the equivalent capacity utility, the proportional fairness utility and the max-min fairness utility, respectively.

that the performance obtained by solving Problem **P2** is independent of the value of θ as long as θ satisfies (9). This is because that by Proposition 1, the optimal solution to Problem **P2** achieves the optimal objective of Problem **P1** when θ satisfies (9).

B. Performance Comparison

To the best of our knowledge, there are no algorithms proposed with the same goal in the literature. For the comparison with our proposed joint optimization of BS association and power control, we therefore introduce two baseline schemes: the max-SNR BS association scheme [16] and the random BS association scheme [45]. The max-SNR scheme works as follows: each MU i is first associated with the BS that poses the strongest SNR when the MU i transmits with the maximum transmit power, and then all MUs optimize their transmit power by Procedure 1. The random scheme works as follows: each MU i is first associated with one BS randomly, and then all MUs optimize their transmit power by Procedure 1.

Example 3 (Performance comparison at different densities of MUs) We consider a set of uplink SCN topologies, where we vary the number of MUs from 20 to 60, and set the number of BSs to be 4. Specifically, the four SBSs are respectively placed in (500m, 500m), (1500m, 500m), (1500m, 1500m) and (500m, 1500m), and all 50 MUs are uniformly distributed in an 1000-meter two-dimensional circle plane whose center is at (1000m, 1000m). The coverage of each SBS is 1000m. Each point in Fig. 6 is obtained by averaging over 100 different topologies with the same density of MUs.

From Fig. 6, it is seen that the obtained utility, the number of served MUs, and the total power consumption all increase with the increase of the number of MUs for any scheme. This is because that more possible MUs can be served with increasing the number of MUs, however, the increase of MUs in service requires more transmit power to combat the more severe co-channel interference. It is not surprising to see that the proposed algorithm always performs better than the baseline schemes. For example, compared with the max-SNR BS association scheme and random BS association scheme, the max-min fairness utility and the proportional fairness utility increase 23% and 30% on average, respectively. This is mainly because that compared to the baseline schemes, the proposed algorithm makes every MU have more freedom to choose its corresponding BS when maximizing the number of served MUs with minimum total power consumption.

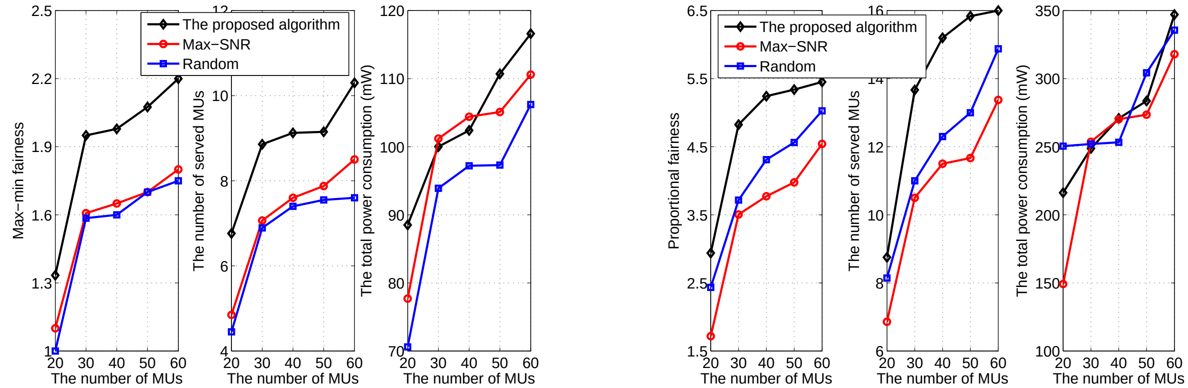
Compared with the proportional fairness utility, the max-min fair utility can guarantee the fairer load balancing among BSs. Comparing Fig. 6(a) with Fig. 6(b), however, we can see that given the number of MUs, the fairer load balancing among BSs leads to fewer MUs in service. It implies that we need to carefully determine the system-wide utility when certain system performance is required.

Example 4 (Performance comparison at different densities of SBSs) We consider a set of uplink SCN topologies, where we vary the number of SBSs from 2 to 6, and set the number of MUs to be 50. Specifically, all BSs and MUs are randomly distributed in an 1000-meter two-dimensional circle plane whose center is at (1000m, 1000m). Each point in Fig. 7 is obtained by averaging over 100 different topologies with the same density of BSs.

It is not surprising to see that the proposed algorithm always performs better than the baseline schemes. From Fig. 7(a), we can find that the max-min fairness utility decreases with increasing the number of SBSs for all three schemes. On the contrary, it is shown in Fig. 7(b) that the proportional fairness utility first increases and then decreases with increasing the number of SBSs for all three schemes. Furthermore, it can be seen that the number of served MUs and the total power consumption both first increase and then decrease with the increase of the number of SBSs for the proposed algorithm and the random BS association scheme. Since all SBSs must be active to serve existing MUs for maximizing the max-min fairness utility and proportional fairness utility and all SBSs use the same spectrum, increasing the density of SBSs results in the severer co-channel interference. When the co-channel interference can be alleviated by the joint optimization of BS

1536-1276 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

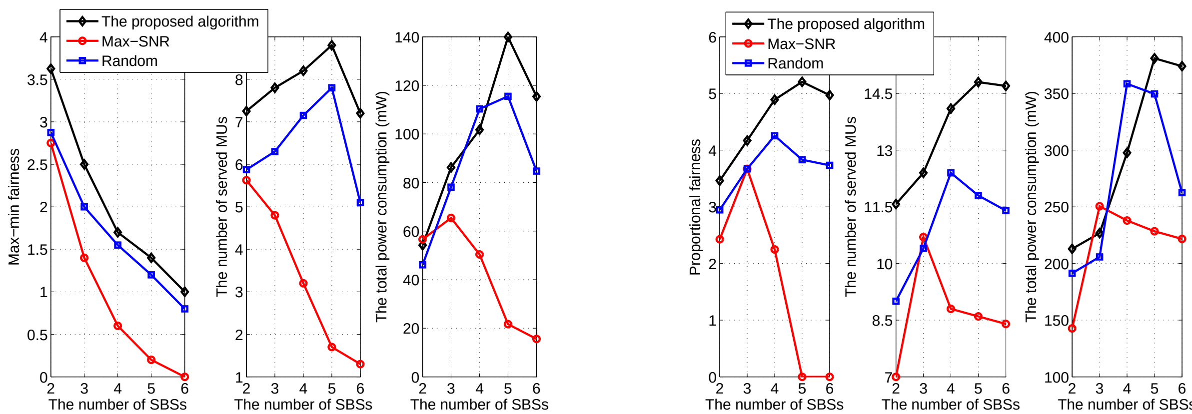
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2017.2664832, IEEE Transactions on Wireless Communications



(a) The max-min fairness

(b) The proportional fairness

Fig. 6. The system-wide utility, the number of served MUs, and the total power consumption vs. the density of MUs.



(a) The max-min fairness

(b) The proportional fairness

Fig. 7. The system-wide utility, the number of served MUs, and the total power consumption vs. the density of SBSs.

association and power control, increasing SBSs can increase the number of served MUs (and hence the system-wide utility and total power consumption). On the contrary, the increase of BSs will lead to the decrease of the number of served MUs (and hence the degradation of system-wide utility and the reduction of total power consumption) when the severe co-channel interference cannot guarantee the minimum average-rate requirements of served MUs any longer. Therefore, although the deployment of SCNs can improve the network performance, the performance improvement is bounded due to the co-channel interference. This observation implies that to serve more MUs, effective spectrum allocation is indispensable along with increasing the density of SBSs.

Example 5 (Performance comparison at different minimum average-rate requirements) We consider a set of uplink SCN topologies, where 4 SBSs and 50 MUs are randomly distributed in an 1000-meter two-dimensional circle plane entered at (1000m, 1000m). In this simulation example, we vary the minimum average-rate requirement from 1.0 Mbps to 5.0 Mbps for all MUs. Fig. 8 shows the obtained system-wide utility, the number of served MUs and the total transmit power averaged over 100 different topologies for three schemes under different system-wide utilities.

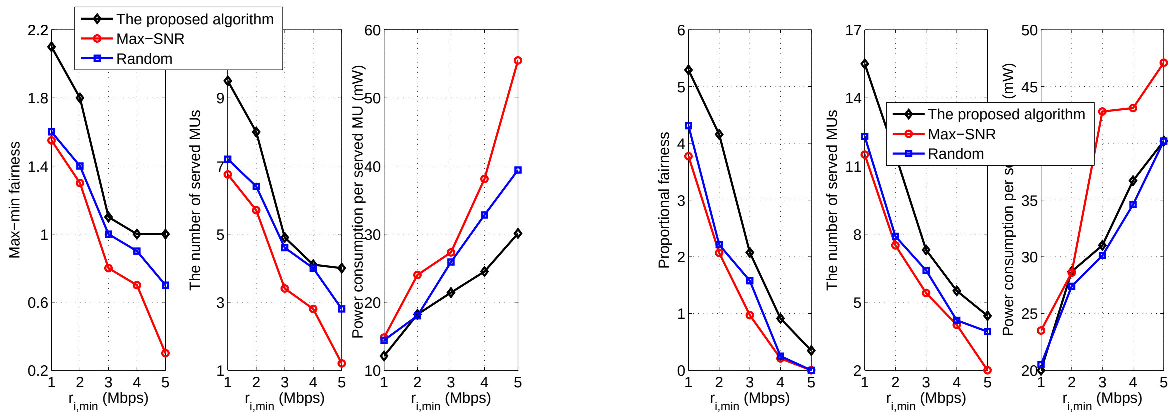
It can be seen from Fig. 8 that the obtained system-wide utility and the number of served MUs both decrease with the increase of the minimum average-rate requirement. This is mainly because that more power consumption is needed for every MUs to satisfy the increase of the minimum average-rate requirement, which leads to the stronger interference among MUs. Furthermore, we find that the power consumption per served MUs increases with the increase of the minimum average-rate requirement. This observation is mainly due to that every served MU needs more transmit power to satisfy the increase of the minimum average-rate requirement.

VI. Conclusions

In this paper, we have studied the joint optimization of BS association and power control that maximizes the system-wide utility and meanwhile minimizes the total transmit power for the maximum utility in the uplink SCN using NOMA with SIC. Specifically, we first transform such a two-stage optimization problem into a single-stage optimization problem. Then, the equivalently transformed optimization problem is efficiently solved by the proposed PCSUM algorithm that is developed based on the theories of coalition formation game and primal decomposition. By comparing with the proposed algorithm through extensive simulations, we have gained deeper understanding of the baseline schemes: the max-SNR scheme

1536-1276 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

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/TWC.2017.2664832, IEEE Transactions on Wireless Communications



(a) The max-min fairness

(b) The proportional fairness

Fig. 8. The system-wide utility, the number of served MUs and the total power consumption vs. the minimum average-rate requirement.

and the random scheme. Simulations further show that the proposed algorithm performs better performance in terms of system-wide utility and total power consumption for the joint optimization of BS association and power control.

For our future work, we will develop distributed algorithms to expedite the convergence and reduce the computational complexity. Furthermore, the mobility of MUs will be taken into account when designing the joint BS association and power control for SCNs. On the other hand, the joint optimization of BS association, power control, and SIC ordering is a challenging work on improving the network performance for the uplink SCN using NOMA with SIC, and thus it will be studied in our future work.

Appendix A
Proof of Theorem 1

Under the logarithmic domain transformation, let $\bar{p}_i = \log(p_i)$ and $\bar{r}_{i\boldsymbol{\pi}_j} = \log(r_{i\boldsymbol{\pi}_j})$. Due to the fact that the function $\log(\log(1+e^{x_1}+\cdots+e^{x_n}))$ is concave, the function at the left-hand side of constraint (16b) can be converted into a concave function (i.e., $\log(\log(\sum_{\boldsymbol{\pi}_j\in\boldsymbol{\Pi}_j:i\in\mathcal{S}_j} q_{\boldsymbol{\pi}_j} e^{\bar{r}_{i\boldsymbol{\pi}_j}}))$) through the logarithmic range transformation. Further, we can rewrite the function at the left-hand side of constraint (16c) as $\log(\log(1+e^{f_{ij}}))$ through the logarithmic domain and range transformation, where $f_{ij} = -\log(\sum_{k\in\mathcal{S}_j:\pi_{kj}>\pi_{ij}} e^{-\bar{p}_i+\bar{p}_k} g_{kj} + \sum_{j'=1,j'\neq j}^{N} \sum_{i'\in\mathcal{S}_{j'}} e^{-\bar{p}_i+\bar{p}_{i'}} g_{i'j} + e^{-\bar{p}_i} n_j) + \log g_{ij}$. Since the function $\log(\log(1+e^{x_1}+\cdots+e^{x_n}))$ is concave and nondecreasing in each argument, the concavity of the function f_{ij} leads to the concavity of $\log(\log(1+e^{f_{ij}}))$. Together with the term at the right-hand side of constraint (16c) being $\bar{r}_{i\boldsymbol{\pi}_j}$, it follows that the constraint (16c) can be transformed into a convex set. Finally, we find that the objective is equal to the convex function $\log(\sum_{\forall i\notin\mathcal{S}_{N+1}} e^{\bar{p}_i})$. In summary, Problem **P4** is now converted into a convex optimization problem. Since Problem **P4** is the equivalent problem of Problem **P3**, Theorem 1 follows.

Appendix B
Proof of Proposition 2

The operations of Procedure 1 include two folds in the worst case: the feasibility checking and the optimal solution search. Since we check the feasibility via the interior-point algorithm, the computational complexity of feasibility checking is equal to $O(\sqrt{5(A+B)})$ [52].

After that, we adopt the primal decomposition with the projected subgradient method to find the optimal solution to Problem **P5**. Recall that a linear programming is solved via the interior-point algorithm at each iteration of Procedure 1, and its computational complexity is $O(A^{0.5})$. On the other hand, an Euclidean projection is performed for solving the master problem at each iteration of Procedure 1, which gives the computational complexity of $O(1)$. Therefore, the computational complexity of Procedure 1 is $O(A^{0.5})$ at each iteration.

Next, we verify how many iterations are needed when Procedure 1 obtains the ϵ-optimal solution. For the natational brevity, we use $\boldsymbol{h}(t)$ to denote the subgradient (i.e., $\text{diag}(\boldsymbol{\lambda}^*(t))(\hat{\boldsymbol{Q}}\hat{\boldsymbol{p}}(t) - \hat{\boldsymbol{u}})$) and use $f(\hat{\boldsymbol{r}}(t))$ to denote the objective function (i.e., $\sum_{j=1}^{N}\sum_{\forall i\in\mathcal{S}_j} \hat{p}_{j_i}(\hat{\boldsymbol{r}}(t))$) in the following proof. Assume that $\hat{\boldsymbol{r}}^*$ is the optimal solution to Problem **P5**, and correspondingly the optimal power allocation is $\hat{\boldsymbol{p}}(\hat{\boldsymbol{r}}^*)$. As shown in (25), the projected subgradient method is used to update $\hat{\boldsymbol{r}}(t)$. It implies that we move closer to every point in $\mathcal{R}$ when we project the point $\hat{\boldsymbol{r}}(t)+\alpha_t\boldsymbol{h}(t)$ onto $\mathcal{R}$. Specifically, let $\boldsymbol{y}(t+1) = \hat{\boldsymbol{r}}(t)+\alpha_t\boldsymbol{h}(t)$, and we have

$$\begin{aligned}
&||\hat{\boldsymbol{r}}(t+1) - \hat{\boldsymbol{r}}^*||_2\\
=&||[\boldsymbol{y}(t+1)]_{\mathcal{R}} - \hat{\boldsymbol{r}}^*||_2\\
\leq&||\boldsymbol{y}(t+1) - \hat{\boldsymbol{r}}^*||_2\\
=&||\hat{\boldsymbol{r}}(t) - \alpha_t\boldsymbol{h}(t) - \hat{\boldsymbol{r}}^*||_2^2\\
=&||\hat{\boldsymbol{r}}(t) - \hat{\boldsymbol{r}}^*||_2^2 + \alpha_t^2||\boldsymbol{h}(t)||_2^2 - 2\alpha_t(\boldsymbol{h}(t))^T(\hat{\boldsymbol{r}}(t) - \hat{\boldsymbol{r}}^*)\\
\leq&||\hat{\boldsymbol{r}}(t) - \hat{\boldsymbol{r}}^*||_2^2 + \alpha_t^2||\boldsymbol{h}(t)||_2^2 - 2\alpha_t(f(\hat{\boldsymbol{r}}(t)) - f(\hat{\boldsymbol{r}}^*))
\end{aligned} \tag{28}$$

where $(\cdot)^T$ denotes transpose. The last line in (28) follows from the definition of subgradient, which gives $f(\hat{\boldsymbol{r}}^*) \geq f(\hat{\boldsymbol{r}}(t)) + (\boldsymbol{h}(t))^T(\hat{\boldsymbol{r}}(t) - \hat{\boldsymbol{r}}^*)$. Applying the inequality (28)

1536-1276 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

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/TWC.2017.2664832, IEEE Transactions on Wireless Communications

recursively, we have

$$||\hat{\boldsymbol{r}}(t+1)-\hat{\boldsymbol{r}}^*||_2 \leq ||\hat{\boldsymbol{r}}(1)-\hat{\boldsymbol{r}}^*||_2^2+\sum_{t_0=1}^{t}\alpha_{t_0}^2||\boldsymbol{h}(t_0)||_2^2 - 2\sum_{t_0=1}^{t}\alpha_{t_0}(f(\hat{\boldsymbol{r}}(t_0))-f(\hat{\boldsymbol{r}}^*)). \quad (29)$$

Due to the fact that $||\hat{\boldsymbol{r}}(t+1)-\hat{\boldsymbol{r}}^*||_2^2 \geq 0$, it follows that

$$2\sum_{t_0=1}^{t}\alpha_{t_0}(f(\hat{\boldsymbol{r}}(t_0))-f(\hat{\boldsymbol{r}}^*)) \leq ||\hat{\boldsymbol{r}}(1)-\hat{\boldsymbol{r}}^*||_2^2+\sum_{t_0=1}^{t}\alpha_{t_0}^2||\boldsymbol{h}(t_0)||_2^2. \quad (30)$$

Together with

$$\sum_{t_0=1}^{t}\alpha_{t_0}(f(\hat{\boldsymbol{r}}(t_0))-f(\hat{\boldsymbol{r}}^*)) \geq (\sum_{t_0=1}^{t}\alpha_{t_0})\min_{t_0=1,\cdots,t}(f(\hat{\boldsymbol{r}}(t))-f(\hat{\boldsymbol{r}}^*)), \quad (31)$$

we have the inequality

$$\min_{t_0=1,\cdots,t} f(\hat{\boldsymbol{r}}(t_0))-f(\hat{\boldsymbol{r}}^*) \leq \frac{||\hat{\boldsymbol{r}}(1)-\hat{\boldsymbol{r}}^*||_2^2+\sum\limits_{t_0=1}^{t}\alpha_{t_0}^2||\boldsymbol{h}(t_0)||_2^2}{2\sum\limits_{t_0=1}^{t}\alpha_{t_0}}. \quad (32)$$

Since $\boldsymbol{\lambda}^*(t)$ is the dual variables of the linear programming (23) and $2^{\hat{r}_{\boldsymbol{b}_{j_i}}(t)}$ is always less than the extreme case of $\frac{g_{j_ij}P_{j_i,\max}+n_j}{n_j}$, there exists a positive number Z that bounds the norm of the subgradient, i.e., $||\boldsymbol{h}(t)||_2 \leq Z$ for all t. By (32), we therefore obtain the following inequality

$$\min_{t_0=1,\cdots,t} f(\hat{\boldsymbol{r}}(t_0))-f(\hat{\boldsymbol{r}}^*) \leq \frac{||\hat{\boldsymbol{r}}(1)-\hat{\boldsymbol{r}}^*||_2^2+Z^2\sum\limits_{t_0=1}^{t}\alpha_{t_0}^2}{2\sum\limits_{t_0=1}^{t}\alpha_{t_0}}. \quad (33)$$

Recall that $||\hat{\boldsymbol{r}}(1)-\hat{\boldsymbol{r}}^*||_2^2$ is less than $\sum\limits_{j=1}^{N}\sum\limits_{\forall i\in\mathcal{S}_j}\log(1+\frac{g_{j_i}P_{j_i,\max}}{n_j})$ which is denoted by C for the brevity, and finally we have

$$\min_{t_0=1,\cdots,t} f(\hat{\boldsymbol{r}}(t_0))-f(\hat{\boldsymbol{r}}^*) \leq \frac{C+Z^2\sum\limits_{t_0=1}^{t}\alpha_{t_0}^2}{2\sum\limits_{t_0=1}^{t}\alpha_{t_0}}. \quad (34)$$

Therefore, if we have

$$\frac{C+Z^2\sum\limits_{t_0=1}^{t}\alpha_{t_0}^2}{2\sum\limits_{t_0=1}^{t}\alpha_{t_0}} \leq \epsilon, \quad (35)$$

the ϵ-optimal solution is obtained. Using the fact that $\alpha_t=\frac{1}{t}$, we have $\sum\limits_{t=1}^{\infty}\alpha_t^2 \to \frac{\pi^2}{6}$ and $\sum\limits_{t_0=1}^{t}\alpha_t > 1+\frac{1}{2}\log(t)$. Therefore, we have the number of iterations needed for the ϵ-optimal solution to be

$$t \geq 2^{\frac{C+Z^2\pi^2/6}{\epsilon}-2} \to O(e^{1/\epsilon}). \quad (36)$$

Since the computational complexity is $O(A^{0.5})$ at each iteration of Procedure 1, it follows that the computational complexity is $O(A^{0.5}e^{1/\epsilon})$ when the ϵ-optimal solution is obtained by Procedure 1. Considering the feasibility checking has the computational complexity of $O(\sqrt{5(A+B)})$ [52], it follows that the total computational complexity of Procedure 1 is of $O(\sqrt{5(A+B)})+O(A^{0.5}e^{1/\epsilon})$ iterations, as shown in Proposition 2.

References

[1] V. Huang and W. Zhuang, “QoS-oriented access control for 4G mobile multimedia CDMA communications,” *IEEE Commun. Mag.*, vol. 40, no. 3, pp. 118-125, Mar. 2002.

[2] M. Soleimanipour, W. Zhuang, and G. H. Freeman, “Optimal resource management in wireless multimedia wideband CDMA systems,” *IEEE Trans. Mobile Computing*, vol. 1, no. 2, pp. 143-160, Apr. 2002.

[3] W. Song and W. Zhuang, “Multi-service load sharing for resource management in the cellular/WLAN integrated network,” *IEEE Trans. Wireless Commun.*, vol. 8, no. 2, pp. 725-735, Feb. 2009.

[4] A. Ghosh, N. Mangalvedhe, R. Ratasuk, etc., “Heterogeneous cellular networks: From theory to practice,” *IEEE Commun. Mag.*, vol. 50, no. 6, pp. 54-64, Jun. 2012.

[5] S. Mukherjee, “Distribution of downlink SINR in heterogeneous cellular networks,” *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 575-585, Apr. 2012.

[6] Y. Mao, Y. Luo, J. Zhang, and K. B. Letaief, “Energy harvesting small cell networks: feasibility, deployment, and operation,” *IEEE Commun. Mag.*, vol. 53, no. 6, pp. 94-101, Jun. 2015.

[7] M. Ismail, W. Zhuang, E. Serpedin, and K. Qaraqe, “A survey on green mobile networking: from the perspectives of network operators and mobile users,” *IEEE Trans. Commun. Survey & Tutorials*, vol. 17, no. 3, pp. 1535-1556, Aug. 2015.

[8] C. Wang, J. Li, and M. Guizani, “Cooperation for spectral and energy efficiency in ultra-dense small cell networks,” *IEEE Wireless Commun.*, vol. 23, no. 1, pp. 64-71, Feb. 2016.

[9] S. Zhang, N. Zhang, S. Zhou, J. Gong, Z. Niu, and X. Shen, “Energy-aware traffic offloading for green heterogeneous networks,” *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1116-1129, May 2016.

[10] N. Zhang, N. Cheng, A. Gamage, K. Zheng, J. W. Mark, and X. Shen, “Cloud assisted HetNets toward 5G wireless networks,” *IEEE Commun. Mag.*, vol. 53, no. 6, pp. 59-65, Jun. 2015.

[11] L. P. Qian, Y. J. Zhang, Y. Wu, and J. Chen, “Joint Base Station Association and Power Control via Benders Decomposition,” *IEEE Trans. Wireless Commun.*, vol. 12, no. 4, pp. 1651-1665, Apr. 2013.

[12] M. Ismail, A. T. Gamage, W. Zhuang, X. Shen, E. Serpedin, and K. Qaraqe, “Uplink decentralized joint bandwidth and power allocation for energy-efficient operation in a heterogeneous wireless medium,” *IEEE Trans. Commun.*, vol. 63, no. 4, pp. 1483-1495, Apr. 2015.

[13] L. P. Qian, Y. Wu, J. Wang, and W. Zhang, “Energy-efficient distributed user scheduling in relay-assisted cellular networks,” *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 4060-4073, Jun. 2016.

[14] D. Tse and P. Viswanath, *Fundamentals of wireless communication*, Cambridge University Press, 2005.

[15] K. Higuchi and A. Benjebbour, “Non-orthogonal multiple access (NO-MA) with successive interference cancellation for future radio access,” *IEICE Trans. Commun.*, vol. 98, no. 3, pp. 403-414, Mar. 2015.

[16] Z. Liu and M. E. Zarki, “SIR-based call admission control for DS-CDMA cellular systems,” *IEEE J. Sel. Areas Commun.*, vol. 12, no. 4, pp. 638-644, May 1994.

[17] Z. J. Haas, J. H. Winter, and D. S. Johnson, “Simulation results of the capacity of cellular systems,” *IEEE Trans. Veh. Technol.*, vol. 46, no. 4, pp. 805-817, Nov. 1999.

[18] R. Yates, “A framework for uplink power control in cellular radio systems,” *IEEE J. Sel. Areas Commun.*, vol. 13, no. 7, pp. 1341-1347, Sep. 1995.

[19] S. V. Hanly, “An algorithm for combined cell-site selection and power control to maximize cellular spread spectrum capacity,” *IEEE J. Sel. Areas Commun.*, vol. 13, pp. 1332-1340, Sep. 1995.

[20] F. Rashid-Farrokhi, K. J. Liu, and L. Tassiulas, “Downlink power control and base station assignment,” *IEEE Commun. Lett.*, vol. 1, no. 4, pp. 102-104, Apr. 1997.

1536-1276 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

[21] J. W. Lee, R. R. Mazumdar, and N. B. Shroff, "Joint resource allocation and base-station assignment for the downlink in CDMA networks," *IEEE/ACM Trans. Networking*, vol. 14, no. 1, pp. 1-14, Feb. 2006.

[22] R. Stridh, M. Bengtsson, and B. Ottersten, "System evaluation of optimal downlink beamforming with congestion control in wireless communication," *IEEE Trans. Wireless Commun.*, vol. 5, no. 4, pp. 743-751, Apr. 2006.

[23] J. T. Wang, "SINR feedback-based integrated base-station assignment, diversity, power control for wireless networks," *IEEE Trans. Veh. Technol.*, vol. 95, no. 1, pp. 473-484, Jan. 2010.

[24] K. Shen and W. Yu, "Distributed pricing-based user association for downlink heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1100-1113, Jun. 2014.

[25] S. Buzzi and H. V. Poor, "Joint receiver and transmitter optimization for energy-efficient CDMA communications," *IEEE J. Sel. Areas Commun.*, vol. 26, pp. 459-472, Apr. 2008.

[26] E. Aryafar, A. Keshavarz-Haddad, M. Wang, and M. Chiang, "RAT selection games in HetNets," *Proc. of IEEE INFOCOM*, pp. 998-1006, Turin, Italy, Jul. 2013.

[27] V. N. Ha and L. B. Le, "Distributed base station association and power control for heterogeneous cellular networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 1, pp. 282-296, Jan. 2014.

[28] R. Sun, M. Hong, and Z. Q. Luo, "Joint downlink base station association and power control for max-min fairness: computation and complexity," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 6, pp. 1040-1054, Jun. 2015.

[29] L. Imhof and R. Mathar, "Capacity regions and optimal power allocation for CDMA cellular radio," *IEEE Trans. Inf. Theory*, vol. 51, no. 6, pp.2001-2019, Jun. 2005.

[30] L. Jiang, S. Parekh, and J. Walrand, "Base station association game in multi-cell wireless networks," *Proc. of WCNC*, pp. 1616-1621, Las Vegas, USA, Mar. 2008.

[31] S. M. Cheng, S. Y. Lien, F. S. Chu, and K. C. Chen, "On exploiting cognitive radio to mitigate interference in macro/femto heterogeneous networks, " *IEEE Trans. Wireless Commun.*, vol. 18, no. 3, pp. 40-47, Mar. 2011.

[32] K. Son, H. Kim, Y. Yi, and B. Krishnamachari, "Base station operation and user association mechanisms for energy-delay tradeoffs in green cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 8, pp.1525-1536, Sept. 2011.

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

[34] Q. Ye, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "On/off macrocells and load balancing in heterogeneous cellular networks," *Proc. of Globecom*, pp. 3814-3819, Altanta GA, USA, Dec. 2013.

[35] H. Dhillon, R. Ganti, and J. Andrews, "Load-aware modeling and analysis of heterogeneous cellular networks, " *IEEE Trans. Wireless Commun.*, vol. 12, no. 4, pp. 1666-1677, Apr. 2013.

[36] H. Zhang, C. Jiang, J. Cheng, and V. C. M. Leung, "Cooperative interference mitigation and handover management for heterogeneous cloud small cell networks," *IEEE Wireless Commun.*, vol. 22, no. 3, pp. 92-99, Jun. 2015.

[37] C. Liu and L. Wang, "Optimal cell load and throughput in green small cell networks with generalized cell association," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1058-1072, May 2016.

[38] P. Xu, Z. Ding, X. Dai, and H. V. Poor, "NOMA: An information theoretic perspective," 2015. [Online]. Available:http://arxiv.org/abs/1504.07751

[39] 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 Processing Lett.*, vol. 21, no. 12, pp. 1501-1505, Dec. 2014.

[40] L. Lei, D. Yuan, C. K. Ho, and S. Sun, "Joint optimization of power and channel allocation with non-orthogonal multiple access for 5G cellular systems," *Proc. of GLOBECOM*, pp. 1-6, San Diego CA, USA, Dec. 2015.

[41] S. Timotheou and I. Krikidis, "Fairness for non-orthogonal multiple access in 5G systems," *IEEE Signal Processing Lett.*, vol. 22, no. 10, pp. 1647-1651, Oct. 2015.

[42] M. Mollanoori and M. Ghaderi, "Uplink scheduling in wireless networks with successive interference cancellation," *IEEE Trans. Mobile Computing*, vol. 13, no. 5, pp. 1132-1144, May 2014.

[43] K. Son, S. Chong, and G. de Veciana "Dynamic association for load balancing and interference avoidance in multi-cell networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 7, pp. 3566-3576, Jul. 2009.

[44] R. Madan, J. Borran, A. Sampath, N. Bhushan, A. Khandekar, and T. Ji, "Cell association and interference coodination in heterogenous LTE-A cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 12, pp. 1479-1489, Dec. 2010.

[45] S. Lee, K. Son, H. Gong, and Y. Yi, "Base station association in wireless cellular networks: An emulation based approach," *IEEE Trans. Wireless Commun.*, vol. 11, no. 8, pp. 2720-2729, Aug. 2012.

[46] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Trans. Networking*, vol. 8, no. 5, pp. 556-567, Oct. 2000.

[47] J. Huang, V. G. Subramanian, R. Agrawal, and R. A. Berry, "Downlink Scheduling and Resource Allocation for OFDM Systems," *IEEE Trans. Wireless Commun.*, vol. 8, no. 1, pp. 288-296, Jan. 2009.

[48] E. Matskani, N. D. Sidiropoulos, Z. Q. Luo, and L. Tassiulas, "Convex approximation techniques for joint multiuser downlink beamforming and admission control," *IEEE Trans. Wireless Commun.*, vol. 7, pp. 2682-2693, Jul. 2008.

[49] W. Saad, Z. Han, M. Debbah, A. Hjorungnes, and T. Basar, "Coalitional game theory for communication networks," *IEEE Signal Processing Mag.*, vol. 26, no. 5, pp. 77-97, 2009.

[50] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 6, no. 6, pp. 721-741, 1984.

[51] B. Hajek, "Cooling schedules for optimal annealing," *Math. Oper. Res.*, vol. 13, no. 2, pp. 311-329, 1988.

[52] Y. Ye, *Interior Point Algorithms: Theory and Analysis*, Wiley, 1997.

[53] 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.

[54] 3GPP, TR36.931, V9.0.0, "LTE; Evolved universal terrestrial radio sccess (E-UTRA); Radio frequency (RF) requirements for LTE Pico Node B," Release 9, May 2011.

[55] R. Zhang, Z. Zhong, M. Wang, X. Shen, and L. Xie, "Equivalent capacity in carrier aggregation-based LTE-A systems: a probabilistic analysis," *IEEE Trans. Wireless Commun.*, vol. 13, no. 11, pp. 6444-6460, Nov. 2014.

[56] D. Liu, Y. Chen, K. K. Chai, T. Zhang, and M. Elkashlan, "Opportunistic user association for multi-service HetNets using Nash bargainning solution," *IEEE Commun. Lett.*, vol. 18, no. 3, pp. 463-466, Mar. 2014.

Li Ping Qian (S'08-M'10-SM'16) received the PhD degree in Information Engineering from the Chinese University of Hong Kong, Hong Kong, in 2010. She is currently an Associate Professor with the College of Information Engineering, Zhejiang University of Technology, China. From 2010 to 2011, she was a Post-Doctoral Research Assistant with the Department of Information Engineering, The Chinese University of Hong Kong. She was a Visiting Student with Princeton University in 2009. She was with Broadband Communications Research Laboratory, University of Waterloo, from 2016 to 2017. Her research interests lie in the areas of wireless communication and networking, cognitive networks, and smart grids, including power control, adaptive resource allocation, cooperative communications, interference cancellation for wireless networks, and demand response management for smart grids.

Dr. Qian serves as an Associate Editor for the IET Communications. She was a co-recipient of the IEEE Marconi Prize Paper Award in wireless communications (the Annual Best Paper Award of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS) in 2011, the Second-class Outstanding Research Award for Zhejiang Provincial Universities, China, in 2012, the Second-class Award of Science and Technology given by Zhejiang Provincial Government in 2015, and the Best Paper Award from ICC2016. She was also a finalist of the Hong Kong Young Scientist Award in 2011, and a recipient of the scholarship under the State Scholarship Fund of China Scholarship Council for Visiting Scholars in 2015, and Zhejiang Provincial Natural Science Foundation for Distinguished Young Scholars in 2015.

1536-1276 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

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/TWC.2017.2664832, IEEE Transactions on Wireless Communications

Yuan Wu (S'08-M'10-SM'16) received the Ph.D degree in Electronic and Computer Engineering from the Hong Kong University of Science and Technology, Hong Kong, in 2010. He is currently an Associate Professor in the College of Information Engineering, Zhejiang University of Technology, Hangzhou, China. During 2010-2011, he was the Postdoctoral Research Associate at the Hong Kong University of Science and Technology. During 2016-2017, he is the visiting scholar at the Broadband Communications Research (BBCR) group, Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research interests focus on radio resource allocations for wireless communications and networks, and smart grid. He is the recipient of the Best Paper Award in IEEE International Conference on Communications (ICC) in 2016.

Haibo Zhou (M'14) received the Ph.D. degree in information and communication engineering from Shanghai Jiao Tong University, Shanghai, China, in 2014. Since 2014, he has been a Post-Doctoral Fellow with the Broadband Communications Research Group, ECE Department, University of Waterloo. His research interests include resource management and protocol design in cognitive radio networks and vehicular networks.

Xuemin (Sherman) Shen (IEEE M'97-SM'02-F'09) received the B.Sc.(1982) degree from Dalian Maritime University (China) and the M.Sc. (1987) and Ph.D. degrees (1990) from Rutgers University, New Jersey (USA), all in electrical engineering. He is a Professor and University Research Chair, Department of Electrical and Computer Engineering, University of Waterloo, Canada. He is also the Associate Chair for Graduate Studies. Dr. Shen's research focuses on resource management in interconnected wireless/wired networks, wireless network security, social networks, smart grid, and vehicular ad hoc and sensor networks. He is an elected member of IEEE ComSoc Board of Governor, and the Chair of Distinguished Lecturers Selection Committee. Dr. Shen served as the Technical Program Committee Chair/Co-Chair for IEEE Globecom'16, Infocom'14, IEEE VTC'10 Fall, and Globecom'07, the Symposia Chair for IEEE ICC'10, the Tutorial Chair for IEEE VTC'11 Spring and IEEE ICC'08, the General Co-Chair for ACM Mobihoc'15, Chinacom'07 and QShine'06, the Chair for IEEE Communications Society Technical Committee on Wireless Communications, and P2P Communications and Networking. He also serves/served as the Editor-in-Chief for IEEE Network, Peer-to-Peer Networking and Application, and IET Communications; an Associate Editor-in-Chief for IEEE Internet of Things Journal, a Founding Area Editor for IEEE Transactions on Wireless Communications; an Associate Editor for IEEE Transactions on Vehicular Technology, Computer Networks, and ACM/Wireless Networks, etc.; and the Guest Editor for IEEE JSAC, IEEE Wireless Communications, IEEE Communications Magazine, and ACM Mobile Networks and Applications, etc. Dr. Shen received the Excellent Graduate Supervision Award in 2006, and the Outstanding Performance Award in 2004, 2007, 2010, and 2014 from the University of Waterloo, the Premier's Research Excellence Award (PREA) in 2003 from the Province of Ontario, Canada, and the Distinguished Performance Award in 2002 and 2007 from the Faculty of Engineering, University of Waterloo. Dr. Shen is a registered Professional Engineer of Ontario, Canada, an IEEE Fellow, an Engineering Institute of Canada Fellow, a Canadian Academy of Engineering Fellow, a Royal Society of Canada Fellow, and a Distinguished Lecturer of IEEE Vehicular Technology Society and Communications Society.

1536-1276 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.