Utility-Optimal Multi-Pattern Reuse in Multi-Cell Networks

Kyuho Son, *Member, IEEE*, Yung Yi, *Member, IEEE*, and Song Chong, *Member, IEEE*

***Abstract*—Achieving sufficient spatial capacity gain through the use of small cells requires careful consideration of inter-cell interference (ICI) management via BS power coordination coupled with user scheduling inside cells. Optimal algorithms are known to be difficult to implement due to high computation and signaling overhead. This study proposes joint pattern-based ICI management and user scheduling algorithms that are practically implementable. The key idea is to decompose the original problem into two sub-problems in which ICI management is run at a slower time scale than user scheduling. We empirically show that even with such a slow tracking of system dynamics at the ICI management part, the decomposed approach achieves a considerable performance increase compared to conventional universal reuse schemes.**

***Index Terms*—Inter-cell interference (ICI), multi-pattern reuse, ICI management, user scheduling, time-scale decomposed algorithm, multi-cell network, network utility maximization (NUM).**

I. Introduction

To achieve high spatial capacity, wireless cellular networks consider a dense deployment of base stations (BSs) that cover small cells. As a consequence, inter-cell interference (ICI) from neighboring BSs becomes a major source of performance degradation, and the portion of users whose capacity is limited by ICI grows. To attain the full potential gain of multi-cell networks, coordinating the transmissions among BSs to manage ICI effectively is essential. The key intuition of BS coordination is that the achievable rates, which depend on the amount of ICI, can be increased by adaptively turning off some of the neighboring BSs. Thus, there are cases in which the increment of achievable rates preponderates the sacrifice of taking away transmission opportunities at neighboring BSs. In particular, this effect of ICI management is very apparent for users at the edges of cells.

A brute-force approach for mitigating ICI involves the use of a system-wide reuse scheme in the time and/or frequency domain. However, this may waste precious radio resources because users at different geographical locations inside cells prefer different reuse schemes. Several schemes, e.g., fractional frequency reuse (FFR) [1] in Mobile WiMAX, have been proposed to accommodate users in different channel conditions with different reuse factors. However, these a-priori hand-crafted schemes are still far from optimal in the sense that they do not adapt to dynamic network environments, e.g., time-varying user loads/locations. In addition, user scheduling working opportunistically based on perceived time-varying channels must be considered in conjunction with ICI management to achieve a high performance gain.

This paper (1) investigates the coupling dynamics of inter-cell ICI management and intra-cell user scheduling and (2) proposes practically implementable joint ICI management and user scheduling algorithms in multi-cell networks. To that end, a *pattern-based* joint optimal algorithm that tracks time-varying channel conditions is initially proposed, where 'pattern' corresponds to a combination of BS ON/OFF activities. It is then demonstrated that the proposed optimal algorithm is difficult to implement due to its high complexity. The key bottleneck lies in the ICI management part, which requires collecting excessive amount of feedback information from all users and also needs complex operations to make decisions on BS coordination at every time slot. To overcome such complexity, the original optimization problem is decomposed into two sub-problems (user scheduling and pattern-based ICI management) and these are solved them with different time scales. The complexity becomes much lower than that of the optimal algorithm, yet sustains high efficiency in ICI management.

The algorithm based on time-scale decomposition stems from a design rationale in which ICI management may not have to track fast dynamics, e.g., a fast fading channel condition. Instead, it may suffice to run the ICI management scheme following only macroscopic network changes, e.g., user loads/locations, and their average channel conditions. In spite of such slow tracking of system dynamics in ICI management, with the proposed decomposed algorithms, it is empirically shown that the performance increase amounts to about 6$\sim$20% (compared to a conventional universal reuse scheme), corresponding to 1/2$\sim$2/3 of the optimal algorithm (which is practically impossible to implement).

Research pertaining to mitigating ICI has recently received much attention [2]–[7]. Optimal binary power control (BPC) for sum rate maximization was considered in [2]. In other studies [3], [4], optimal joint ICI management (similar BPC) and user scheduling algorithms that operate in a slot-by-slot manner and require heavy computation overhead were considered. The authors there presented the idea of using clustering only with neighboring BSs [3] or considering only neighboring BSs [4] to reduce complexity. However, these

Manuscript received December 5, 2009; revised May 16, 2010 and September 7, 2010; accepted September 19, 2010. The associate editor coordinating the review of this paper and approving it for publication was E. Hossain.

K. Son is with the Department of Electrical Engineering, University of Southern California, Los Angeles, CA, 90089 (e-mail: kyuho.son@usc.edu).

Y. Yi and S. Chong are with the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 305-701, Korea (e-mail: yiyung@kaist.edu; songchong@kaist.edu).

This work was supported by the IT R&D program of MKE/KEIT [KI002137, Ultra Small Cell Based Autonomic Wireless Network].

Digital Object Identifier 10.1109/TWC.2010.110310.091778

schemes continue to require centralized coordination and/or complex operations on a per-slot basis, which hinders practical implementation.

Several recent approaches [6], [7] have attempted to make algorithms practical based on a slightly different time-scale separation approach from that used in the this paper. In one recent study [6], the authors abstract users that share similar traffic loads and channel environments into classes and perform ICI management on a very long time scale (e.g., hours) without explicit consideration of intra-cell user scheduling. They basically design an ICI management scheme that tracks system dynamics at a highly macroscopic level. Our approach differs from the aforementioned scheme [6] in that user scheduling is explicitly considered. Moreover, our ICI management runs much faster (e.g., in the order of seconds) compared to the earlier work [6]. The work with a time-scale separation similar to ours was recently proposed in [7] for different systems, i.e., OFDMA systems, where the transmit power level for different subbands for ICI management is periodically updated. Due to the differences in the system model, a different mechanism is used here that updates patterns and not powers, leading to a different style for the algorithms and analysis. Additionally, the performance gap between optimal and decomposed algorithms is studied in this paper.

Related work also includes an examination of the potential capacity gains (from the perspective of the flow-level performance) by BS coordination [8]. Another important issue in multi-cell networks is to resolve load imbalance problem between cells. Several investigations [9], [10] *explicitly* balance the load by changing user associations from the BS in hot-spot cells to an adjacent BS that is less crowded. Sang et al. [9] proposed an integrated framework consisting of a MAC-layer cell breathing technique and load-aware handover/cell-site selection to deal with load balancing. Bu et al. [10] were the first to consider the formulation of network-wide proportional fairness (PF) [11] rigorously in a multi-cell network where associations between users and BSs are decision variables. Although it is assumed in this paper that user association is fixed, we later argue and empirically show that ICI management can *implicitly* resolve the load imbalance. Moreover, it is shown that any performance gain by controlling user associations may not be significant.

The remainder of this paper is organized as follows. Section II formally describes our system model and presents a definition of the problem. Section III introduces a joint optimal pattern selection and user scheduling algorithm to solve this problem and discuss its implementation difficulties. In order to take into account practical concerns, two algorithms (pattern portion change algorithm and user scheduling algorithm) using time-scale decomposition schemes that run at different time-scales are designed in Section IV. Section V includes a demonstration of the performance of proposed algorithms, and the paper is concluded with Section VI.

II. System Model and Problem Definition

A. Network Model

We consider a wireless cellular network consisting of multiple cells. Denote by $\mathcal{N} \doteq \{1, \ldots, N\}$ and $\mathcal{K} \doteq \{1, \ldots, K\}$ a set of BSs and MSs (or users), respectively. A user $k \in \mathcal{K}$ is associated with a single BS $n \in \mathcal{N}$, which means that data intended for the user k is served only by the BS n. Define $a(\cdot) : \mathcal{K} \to \mathcal{N}$ to be the association function, e.g., $a(k) = n$ if the user k is associated with the BS n. We further denote by $\mathcal{K}_n$ the set of users associated with the BS n.

We assume that a BS transmits data with either its given maximum power or 0, which we simply denote by 'ON' or 'OFF' states[1]. We assume that a same frequency band (or channel in short) with bandwidth W in all cells, and consider only downlink transmissions in the time-slotted system indexed by $t = 0, 1, \ldots$. At each slot, a BS can select only one user for its data transmission. Channels may be time-varying, modeled by some stationary, ergodic random process with the finite state index set $\mathcal{I}$ and the stationary distribution $\boldsymbol{\theta} = (\theta^{(i)},\ i \in \mathcal{I})$.

B. Network Resource and Allocation Schemes

The time-varying network resources at slot t are represented by a finite set $\mathcal{R}(t)$ of the K-dimensional feasible rate (bits/slot) vectors over users. A resource allocation scheme then chooses a feasible rate vector in $\mathcal{R}(t)$ at each slot and serves a subset of users with the chosen rate vector. A feasible rate vector in $\mathcal{R}(t)$ is determined by the following two factors: (i) which BSs are activated and (ii) which users are selected in cells for data transmission.

To formally discuss (i), we define *reuse pattern* (or simply *pattern*) p to be a combination of ON/OFF activities of BSs, which determines inter-cell interference to the corresponding scheduled users in cells. Denote by $\mathcal{P}$ the set of all patterns. A pattern p is said to *activate* a BS n, if the activity of the BS n is ON under pattern p. Denote by $\mathcal{N}_p \subset \mathcal{N}$ the set of all BSs activated by the pattern p. In parallel, we denote by $\mathcal{P}_n \subset \mathcal{P}$ the set of patterns that activate the BS n. Define *reuse factor* of a pattern p to be $\chi_p \doteq \frac{|\mathcal{N}_p|}{N} \leq 1$, i.e., the ratio of the number of BSs which use a pattern p to the total number of BSs. Denote by $X_p(t)$ the *pattern selection indicator* for the pattern p, i.e., $X_p(t) = 1$ when the pattern p is used at slot t, and 0 otherwise. Then, since only one pattern is used per one slot, we should have:

$$\sum_{p \in \mathcal{P}} X_p(t) = 1. \tag{1}$$

In regard to (ii), define *user scheduling indicator* at slot t by $I_k(t)$, i.e., $I_k(t) = 1$, when the user k is scheduled in its associating cell, and 0 otherwise. Reflecting the constraint that only one user can be selected in each cell, we should have:

$$\sum_{k \in \mathcal{K}_n} I_k(t) \begin{cases} \leq 1, & \text{if } X_p(t) = 1 \text{ and } n \in \mathcal{N}_p\ , \\ = 0, & \text{otherwise.} \end{cases} \tag{2}$$

Then, a resource allocation scheme incorporates *pattern selection* and *user scheduling* that can be regarded as choosing a sequence of $\left\{(X_p(t) : p \in \mathcal{P}), (I_k(t) : k \in \mathcal{K})\right\}_{t=0}^{\infty}$ satisfying the constraints (1) and (2).

We now define the transmission rates of users provided by a resource allocation scheme, depending on the choice of

[1]All discussions in this paper can be readily extended to the case of a finite number of discrete power levels.

patterns. Let $G_{n,k}(t)$ represent the time-varying channel gain from BS n to user k at slot t. The channel gain may take into account path loss, log-normal shadowing, fast fading and etc. The received SINR for user k at slot t when pattern p is selected and user k is served by its serving BS, can be written as:

$$\Gamma_{kp}(t)=\begin{cases}\dfrac{G_{a(k),k}(t)P_n^{max}}{N_0W+\sum_{m\in\mathcal{N}_p,m\neq a(k)}G_{m,k}(t)P_m^{max}}, & \text{if } a(k)\in\mathcal{N}_p,\\ 0, & \text{otherwise,}\end{cases}$$

where P_n^{max} is the maximum transmit power of BS n and N_0 is the noise spectral density. Following the Shannon's formula, the data rate for user k on reuse pattern p at slot t is given by:

$$r_{kp}(t)=W\log_2\left(1+\Gamma_{kp}(t)\right).$$

Note that $r_{kp}(t)=0$ for all $a(k)\notin\mathcal{N}_p$, i.e., user k cannot receive any data rate if its serving BS $a(k)$ is not activated by the pattern p. Also notice that $r_{kp}(t)$ is the *potential data rate* when the user k is scheduled, i.e., its actual data rate may become 0, when other user, say k', associated with the BS $a(k)$, is scheduled for service. We assume that each BS n knows instantaneous achievable data rates for all its associated users through channel feedbacks. We further assume that BSs have infinite amount of data to be destined to users.

C. General Problem Statement

In this paper, we aim at proposing the joint pattern selection and user scheduling that maximizes the long-term network-wide utility whenever possible, i.e., solves the following optimization problem **Q**:

$$\textbf{Q:}\quad \max \quad U=\sum_{n\in\mathcal{N}}U^{(n)}=\sum_{k\in\mathcal{K}}U_k(\bar{R}_k)$$
$$\text{subject to}\quad \bar{\mathbf{R}}\in\mathcal{R},$$

where $\bar{\mathbf{R}}=(\bar{R}_k,k\in\mathcal{K})$ is the vector of long-term user throughputs. The network-wide utility U is just the summation of utilities of all BSs $(U^{(n)},n\in\mathcal{N})$; $U^{(n)}$ is again the summation of utilities of all its associated users $U^{(n)}=\sum_{k\in\mathcal{K}_n}U_k(\bar{R}_k)$. Assume the standard condition of differentiability and strictly increasing concavity of $U_k(\cdot)$. We adopt the generalized (w,α)-fair utility function introduced in [12]:

$$U_k(\bar{R}_k)=\begin{cases}w_k\log\bar{R}_k, & \text{if } \alpha=1,\\ w_k(1-\alpha)^{-1}\bar{R}_k^{1-\alpha}, & \text{otherwise,}\end{cases}\tag{3}$$

where α and w_k are positive. By varying the α parameter, it encompasses various notions of fairness, in particular, proportional fairness ($\alpha=1$) and max-min fairness ($\alpha\to\infty$).

The set $\mathcal{R}\subset\mathbb{R}_+^K$ of all achievable rates of users is referred to as *achievable rate region*. First, denote by $\mathcal{R}^{(i)}$ the achievable rate when the system is in the i-th channel state. The $\mathcal{R}^{(i)}$ is essentially the convex hull of the set of feasible rates for the i-th channel state, i.e.,

$$\mathcal{R}^{(i)}=\Big\{\bar{\mathbf{R}}^{(i)}=(\bar{R}_k^{(i)}:k\in\mathcal{K})\mid \exists\,\boldsymbol{\pi}^{(i)}\in\Pi,\ \bar{R}_k^{(i)}=\sum_{p\in\mathcal{P}}\pi_{kp}^{(i)}r_{kp}^{(i)}\Big\},\tag{4}$$

where $\pi_{kp}^{(i)}$ is the long-term fraction of time that user k is served under pattern p for the i-th channel state; Π is the set of nonnegative vectors $\boldsymbol{\pi}^{(i)}=(\pi_{kp}^{(i)}:k\in\mathcal{K},p\in\mathcal{P})$ such that

$$\sum_{p\in\mathcal{P}}\pi_p^{(i)}=1\ \text{ and }\ \sum_{k\in\mathcal{K}_n}\pi_{kp}^{(i)}\le\pi_p^{(i)},\quad\forall n\in\mathcal{N},\forall p\in\mathcal{P}_n.\tag{5}$$

Then, the $\mathcal{R}$ is characterized by $\mathcal{R}=\sum_{i\in\mathcal{I}}\theta^{(i)}\mathcal{R}^{(i)}$, where the addition of sets is defined as follows: $\mathcal{X}+\mathcal{Y}=\{x+y:x\in\mathcal{X},y\in\mathcal{Y}\}$. Thus, we can characterize the achievable rate region $\mathcal{R}$ by:

$$\mathcal{R}=\Big\{\bar{\mathbf{R}}=(\bar{R}_k:k\in\mathcal{K})\mid\exists\,\bar{\mathbf{R}}^{(i)}\in\mathcal{R}^{(i)},\ \bar{R}_k=\sum_{i\in\mathcal{I}}\theta^{(i)}\bar{R}_k^{(i)}=\sum_{i\in\mathcal{I}}\sum_{p\in\mathcal{P}}\theta^{(i)}\pi_{kp}^{(i)}r_{kp}^{(i)}\Big\}.\tag{6}$$

III. Optimal Algorithm

In this section, the structure of optimal solutions is studied analytically for a simple scenario to gain insight and an optimal pattern selection and user scheduling algorithm that generates the optimal solution is described.

A. Structure of Optimal Solution for Symmetric Networks with Static Channels

For general networks, it is quite difficult to characterize the optimal fractions of time for user-patterns $(\pi_{kp}^{(i)}:k\in\mathcal{K},p\in\mathcal{P},i\in\mathcal{I})$. However, we will show that it is indeed possible to explicitly characterize them for symmetric networks with static channels. A network is said to be symmetric if all BSs have the same number of users, and their channel characteristics are identical. Fig. 1 depicts an illustrative example of a linear two-cell network having three patterns where $(\chi_1,\chi_2,\chi_3)=(1,0.5,0.5)$. Recall that χ_p, the reuse factor of pattern p, is the ratio of the number of BSs activated by pattern p to the total number of BSs. Since the network is symmetric, it is enough to analyze the following optimization problem **Q-symmetric** for a typical BS, say BS 1:

Q-symmetric:

$$\max_{(\pi_{kp}:k\in\mathcal{K}_1,p\in\mathcal{P}_1)}\quad U^{(1)}=\sum_{k\in\mathcal{K}_1}U_k(\bar{R}_k)\tag{7}$$
$$\text{subject to}\quad \sum_{p\in\mathcal{P}_1}\sum_{k\in\mathcal{K}_1}\frac{\pi_{kp}}{\chi_p}\le1,\tag{8}$$
$$\pi_{kp}\ge0,\quad\forall k\in\mathcal{K}_1,\forall p\in\mathcal{P}_1,\tag{9}$$
$$\bar{R}_k=\sum_{p\in\mathcal{P}_1}\pi_{kp}r_{kp},\quad\forall k\in\mathcal{K}_1.\tag{10}$$

Here, the constraint (8) originally comes from the condition (5) on π_{kp}. By additionally applying the symmetric condition that the pattern having the same reuse factor should have the same pattern portion to (5), we can readily derive the constraint (8). For example, in the two-cell network case, the derivation can be done as follows:

$$\begin{aligned}1&=\textstyle\sum_{p\in\mathcal{P}}\pi_p=\pi_1+\pi_2+\pi_3\\&=\pi_1+2\pi_2=\pi_1/\chi_1+\pi_2/\chi_2\ \ (\because\pi_2=\pi_3\text{ by symmetry})\\&\ge\textstyle\sum_{k\in\mathcal{K}_1}\frac{\pi_{k1}}{\chi_1}+\sum_{k\in\mathcal{K}_1}\frac{\pi_{k2}}{\chi_2}=\sum_{p\in\mathcal{P}_1}\sum_{k\in\mathcal{K}_1}\frac{\pi_{kp}}{\chi_p}.\end{aligned}$$

The problem **Q-symmetric** has an interesting structure of optimal solution stated by Lemmas 3.1 and 3.2. Let $\chi_p r_{kp}$

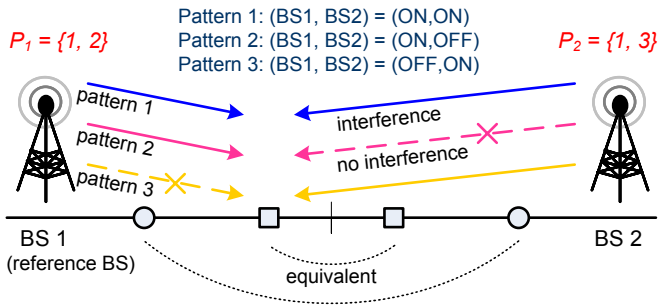


Fig. 1. Example of a linear two-cell network.

be the *effective rate* on pattern p for user k, which is the normalized data rate w.r.t. χ_p. Note that there is a trade-off between the reuse factor χ_p and the data rate r_{kp}. If the user k chooses the pattern p with the lower value χ_p, then the less BSs are active in the network, and accordingly the higher data rate r_{kp} is expected, and vice versa.

Lemma 3.1: For symmetric networks with static channels, the objective (7) is maximized if and only if

$$\pi_{kp} \begin{cases} \geq 0, & \text{if } p = p^*(k), \\ = 0, & \text{otherwise,} \end{cases} \quad \text{where } p^*(k) = \arg\max_p \chi_p r_{kp}.$$

This implies that each user, if served, only utilizes the pattern having the largest effective rate. For simplicity, we ignore the case when more than two patterns have the same largest value throughout the analysis in Section III. Accordingly, each user k can always have the only one optimal pattern $p = p^*(k)$ with $\pi_{kp} > 0$.

Lemma 3.2: For the generalized (w, α)-fair utility function, the optimal fractions of time for user-patterns is given by:

$$\pi_{kp^*(k)} = \left(w_k \chi_{p^*(k)} r_{kp^*(k)}^{1-\alpha} / \lambda_0\right)^{1/\alpha} \text{ and}$$
$$\lambda_0 = \left(\sum_{p\in\mathcal{P}_1} \sum_{k\in\mathcal{K}_{1p}} w_k^{1/\alpha} \chi_{p^*(k)}^{\frac{1-\alpha}{\alpha}} r_{kp^*(k)}^{\frac{1-\alpha}{\alpha}}\right)^{\alpha}, \tag{11}$$

where $\mathcal{K}_{1p}$ is the set of users whose most effective pattern having the highest effective rate is p, i.e., $p = p^*(k) = \arg\max_p \chi_p r_{kp}$ if $k \in \mathcal{K}_{1p}$.

Please refer to Appendix for proofs of these lemmas. Now, we give a numerical example to illustrate the property of the optimal solution.

B. Example: A Linear Two-Cell Symmetric Network

Consider the example of the linear two-cell symmetric network in the Fig. 1. In this example, we have three patterns $p \in \mathcal{P} = \{1, 2, 3\}$ where $\mathcal{N}_1 = \{1, 2\}, \mathcal{N}_2 = \{1\}, \mathcal{N}_3 = \{2\}$ and $(\chi_1, \chi_2, \chi_3) = (1, 0.5, 0.5)$. Suppose that all users have the same utility function with $(w, \alpha) = (1, 1)$. Then, we can obtain $\lambda_0 = |\mathcal{K}_1|$ from (11) regardless of the values of $r_{kp*(k)}$. Let us denote by $\mathcal{K}_{11}$ and $\mathcal{K}_{12}$ the set of users such that $r_{k1} \geq 2r_{k2}$ (i.e., the set of center users) and the set of users such that $r_{k1} < 2r_{k2}$ (i.e., the set of edge users), respectively. Accordingly, the optimal pattern for each user $k \in \mathcal{K}_1$, i.e., $p^*(k) = \arg\max_p \chi_p r_{kp}$, is equal to 1 if $k \in \mathcal{K}_{11}$, and 2 otherwise. Thus, we can obtain the optimal

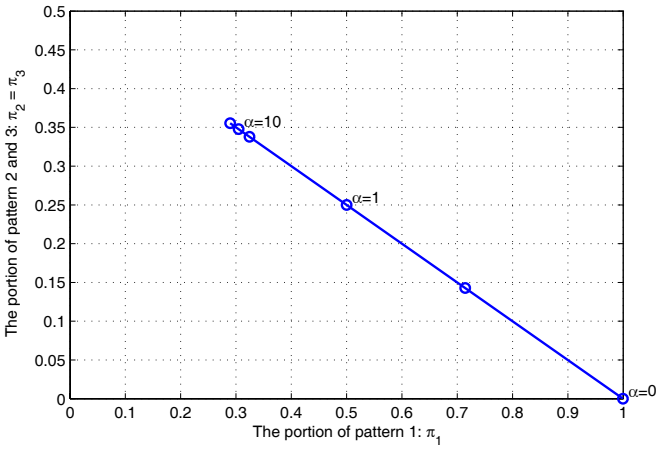


Fig. 2. Numerical example of the linear two-cell symmetric network where each BS has two users; user 1 is in the inner region of the cell and user 2 is in the edge of the cell, whose instantaneous data rate vectors are given by $(r_{11}, r_{12}) = (10, 11)$ and $(r_{21}, r_{22}) = (3, 8)$.

time fractions of user-patterns and the optimal portion for each pattern (π_1, π_2, π_3) as follows:

$$\pi_{kp^*(k)} = \begin{cases} |\mathcal{K}_1|^{-1}, & \text{if } k \in \mathcal{K}_{11}, \\ (2|\mathcal{K}_1|)^{-1}, & \text{if } k \in \mathcal{K}_{12}, \end{cases} \tag{12}$$

$$\pi_1 = \sum_{k\in\mathcal{K}_{11}} \pi_{kp^*(k)} = |\mathcal{K}_{11}|/|\mathcal{K}_1| \quad \text{and}$$
$$\pi_2 = \pi_3 = \sum_{k\in\mathcal{K}_{12}} \pi_{kp^*(k)} = |\mathcal{K}_{12}|/(2|\mathcal{K}_1|)\,. \tag{13}$$

Note that in the case of proportional fair ($\alpha = 1$) the optimal portion of each pattern depends only on and is proportional to the number of users in the sets of center and edge users. However, for general cases ($\alpha \neq 1$), its closed form is very complex because the optimal portion of each pattern depends on the data rate $r_{kp^*(k)}$ for all users due to (11). Thus, we rely on numerical computations for $\alpha \neq 1$.

Fig. 2 depicts the optimal portion of patterns with respect to the fairness criterion α. We fix the number of users as shown in the Fig. 1, that is, each BS has two users: one is in the center and the other in the edge of the cell, $|\mathcal{K}_1| = 2, |\mathcal{K}_{11}| = 1, |\mathcal{K}_{12}| = 1$. When $\alpha = 1$, the optimal portion of patterns can be given by (13): $(\pi_1, \pi_2, \pi_3) = (1/2, 1/4, 1/4)$. Accordingly, user throughputs can be easily calculated: $(\bar{R}_1, \bar{R}_2) = (\pi_{11} r_{11}, \pi_{22} r_{22}) = (\pi_1 r_{11}, \pi_2 r_{22}) = (5, 2)$. When we increase α, i.e., enforcing more fairness, the portions of pattern 2 and 3 avoiding ICI increase in order to increase the throughput of edge users. On the other hand, when we decrease α, the portion of pattern 1 increases as expected. In the extreme case, throughput maximization (α goes 0), only user 1 having a better channel is always served with pattern 1, and user 2 cannot be served at all, i.e., $(\pi_1, \pi_2, \pi_3) = (1, 0, 0)$.

C. Joint Optimal Pattern Selection and User Scheduling Algorithm

We now present an joint optimal pattern selection and user scheduling algorithm. To that end, we use a stochastic gradient-based algorithm, e.g., [13] (only considering user scheduling in a single-cell system), that selects the achievable rate vector maximizing the sum of weighted rates where the

weights are marginal utilities at each slot. Then, it suffices to solve the following problem at each slot, which jointly determines the pattern selection $\mathbf{X}(t) = (X_p(t) : p \in \mathcal{P})$ and user scheduling $\mathbf{I}(t) = (I_k(t) : k \in \mathcal{K})$:

Q-joint: (14)

$$\max_{\mathbf{X}(t),\mathbf{I}(t)} \quad \Delta U(t) = \sum_{k\in\mathcal{K}} U_k'(\bar{R}_k(t-1))r_k(t) \tag{15}$$

$$\text{subject to} \quad \sum_{p\in\mathcal{P}} X_p(t) = 1, \tag{16}$$

$$\sum_{k\in\mathcal{K}_n} I_k(t) \begin{cases} \leq 1, & \text{if } X_p(t) = 1 \text{ and } n \in \mathcal{N}_p, \\ = 0, & \text{otherwise,} \end{cases} \tag{17}$$

where $r_k(t) = \sum_{p\in\mathcal{P}} X_p(t)I_k(t)r_{kp}(t)$ is the actual data rate assigned to user k at slot t and $\bar{R}_k(t) = \frac{1}{t}\sum_{\tau=1}^{t} r_k(\tau) = \bar{R}_k(t-1) + \epsilon_t\left[r_k(t) - \bar{R}_k(t-1)\right]$ (by letting $\epsilon_t = 1/t$) is the long-term throughput for user k up to slot t.

Remark 3.3: If we fix the user scheduling $\mathbf{I}(t)$ and choose utility function as $U_k(\bar{R}_k) = \bar{R}_k$ in **Q-joint**, then this problem is reduced to binary power control (BPC) problem for sum-rate maximization in [2].

The problem **Q-joint** can be naively solved by an exhaustive search. For each pattern p, it needs to compare all possible combinations of user scheduling for all BSs. Thus, this naive approach requires $O(P \cdot K^N)$ complexity, which is computationally intractable. However, Lemma 3.4 tells us the nice property of the problem that we need to consider only the case with the best users selected by intra-cell user scheduling in (18) instead of all possible combination of user scheduling for each pattern.

Lemma 3.4: The problem **Q-joint** can be decomposed into the following $|\mathcal{N}_p|$ independent intra-cell user scheduling problems for a *given* pattern p:

$$k_n^*(t) = \arg\max_{k\in\mathcal{K}_n} U_k'(\bar{R}_k(t-1))r_{kp}(t), \quad \forall n \in \mathcal{N}_p. \tag{18}$$

Please refer to Appendix for a proof.

With the help of Lemma 3.4, we can develop the joint optimal pattern selection and user scheduling algorithm. For each pattern p, we select the best user having the largest value of $U_k'(\bar{R}_k(t-1))r_{kp}(t)$ from (18) and then the best pattern $p^*(t)$ that maximizes the sum of weighted rate $U_k'(\bar{R}_k(t-1))r_{kp^*}(t)$ of the scheduled users. Note that it has much lower complexity[2] $O(P \cdot \sum_{n\in\mathcal{N}} K_n) = O(P \cdot K)$ than that of exhaustive search $O(P \cdot K^N)$. The proof of convergence to the optimal solution is a slight extension to [13], [14] that studied only user scheduling for a fixed pattern. We skip the proof.

Joint pattern selection and user scheduling algorithm

At each slot t, compute $(p^*(t), k_n^*(t), n \in \mathcal{N})$ satisfying

$$\begin{aligned} p^*(t) &= \arg\max_{p\in\mathcal{P}} \sum_{n\in\mathcal{N}_p} \left[\max_{k\in\mathcal{K}_n} U_k'(\bar{R}_k(t-1))r_{kp}(t)\right], \\ k_n^*(t) &= \arg\max_{k\in\mathcal{K}_n} U_k'(\bar{R}_k(t-1))r_{kp^*}(t), \quad \forall n \in \mathcal{N}_{p^*}. \end{aligned} \tag{19}$$

[2]The maximum operation in the intra-cell user scheduling requires linear complexity in the number of users.

This joint optimal pattern selection and user scheduling algorithm requires instantaneous channel feedbacks from all users in the network. We assume that at each slot t, user k estimates its own SINR for all patterns $p \in \mathcal{P}_{a(k)}$ upon listening to pilot signals, calculates the instantaneous data rate $r_{kp}(t)$ and then reports this information to the cental coordinator through its serving BS.

However, this joint optimal algorithm still has implementation difficulties. Apart from the computational complexity of this algorithm, the central coordinator running the algorithm needs to collect the following information from each BS $n \in \mathcal{N}$: instantaneous data rate $r_{kp}(t)$ of all its associated users $k \in \mathcal{K}_n$ on its available patterns $p \in \mathcal{P}_n$. The total amount of feedbacks is quite large, i.e, $(\sum_{n\in\mathcal{N}} |\mathcal{K}_n||\mathcal{P}_n|)$, though they may be delivered along with high speed wired links. Furthermore, a series of tasks, including information feedback from BSs to the central coordinator as well as the computation and the distribution of central coordinator's decision, should be performed in one slot.

IV. Time-Scale Decomposed Algorithm

A. Algorithm Description

In contrast to the centralized joint pattern selection and user scheduling algorithm in Section III, user scheduling in practice is typically undertaken by individual BSs independently without any coordination or information exchanges with other BSs. In this section, in order to take into account such autonomous features in user scheduling and to overcome high computation and feedback overhead in the optimal algorithm, user scheduling is run at every slot, whereas pattern portion change less frequently, in this case, every $T_p >> 1$ slots. We first describe the proposed algorithm (see Fig. 3 for a pictorial description) and then explain the rationale behind it.

Pattern portion change algorithm

Initialization: $\pi_p = 1/|\mathcal{P}|$ for all $p \in \mathcal{P}$.

For every T_p slots, each BS $n \in \mathcal{N}$ computes the partial derivative $D_p^{(n)} \doteq \partial U^{(n)}/\partial\pi_p$ and sends it to the central coordinator,

$$D_p^{(n)} = \sum_{k\in\mathcal{K}_n} U_k'(\bar{R}_k)\cdot\left(\frac{\bar{\pi}_{kp}}{\pi_p}\bar{r}_{kp}\right), \quad p \in \mathcal{P}_n. \tag{20}$$

Then, the central coordinator calculates the gradient vector $\mathbf{D} = (D_1, D_2, \cdots, D_P)$ by collecting $D_p^{(n)}$ from all BSs,

$$D_p = \sum_{n\in\mathcal{N}} D_p^{(n)}, \quad p \in \mathcal{P}, \tag{21}$$

and updates the pattern portion vector $\boldsymbol{\pi}$ as follows,

$$\boldsymbol{\pi} \leftarrow Proj_{\sum_{p\in\mathcal{P}}\pi_p=1}(\boldsymbol{\pi} + \gamma\mathbf{D}), \tag{22}$$

where $Proj_A(\cdot)$ denotes an orthogonal projection on a set A.

User scheduling algorithm

Initialization: $\bar{R}_k(0) = \bar{\pi}_{kp}(0) = \bar{r}_{kp}(0) = 0$.

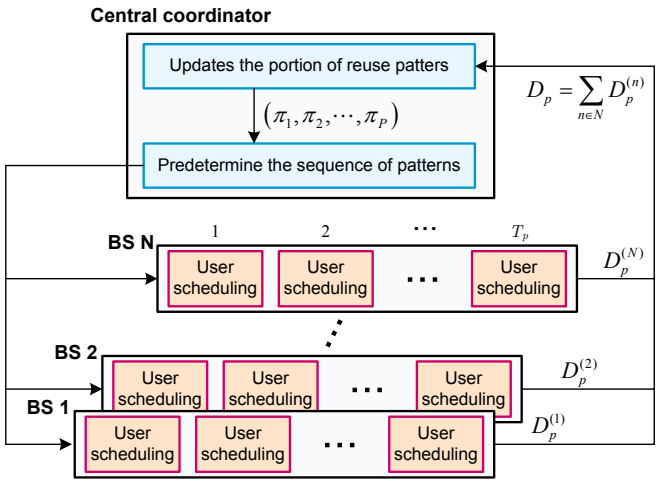


Fig. 3. Proposed time-scale decomposed algorithms.

At each slot t, each BS $n \in \mathcal{N}_{p(t)}$ activated by pattern $p(t)$ selects the user $k_n^*(t)$, i.e., $I_{k_n^* p}(t) = 1$,

$$k_n^*(t) = \arg\max_{k \in \mathcal{K}_n} U_k'(\bar{R}_k(t-1)) r_{kp}(t), \text{[3]} \tag{23}$$

and updates the following variables for all users $k \in \mathcal{K}_n$ with some constants $0 < \beta_1, \beta_2, \beta_3 < 1$:

$$\begin{aligned} \bar{R}_k(t) &= (1-\beta_1)\bar{R}_k(t-1) + \beta_1 I_k(t) r_{kp}(t) \,, \\ \bar{\pi}_{kp}(t) &= (1-\beta_2)\bar{\pi}_{kp}(t-1) + \beta_2 I_k(t) \,, \\ \bar{r}_{kp}(t) &= \begin{cases} (1-\beta_3)\bar{r}_{kp}(t-1) + \beta_3 r_{kp}(t), & \text{if } I_k(t) = 1 \,, \\ \bar{r}_{kp}(t-1), & \text{otherwise.} \end{cases} \end{aligned} \tag{24}$$

Two algorithms with different time scales interact with each other as follows: The pattern portion change algorithm adjusts the portion of reuse patterns $\boldsymbol{\pi}$ for every T_p slots, using the variables $\bar{R}_k(t), \bar{\pi}_{kp}(t)$ and $\bar{r}_{kp}(t)$. These variables essentially correspond to the long-term averages of $I_k(t) r_{kp}(t)$, $\pi_{kp}(t)$, and $r_{kp}(t)$ which are progressively updated at every slot by the user scheduling algorithm. This time-scale decomposition and the way of interaction between the two algorithms implies that the pattern portion algorithm is designed and operated so that it tracks only the *average* interference levels and channel conditions rather than instantaneous conditions such as the joint optimal algorithm in Section III. Considering that a user scheduling algorithm can be carried out autonomously, the actual (amortized) complexity and message passing overhead per slot can be significantly reduced, which makes the proposed algorithms much more implementable. Subsection IV-C includes a discussion of the value of such complexity reduction, i.e., a utility performance gap compared to the optimal algorithm.

B. Rationale of Time-scale Decomposed Algorithms

The pattern portion change algorithm can be regarded as a standard gradient projection algorithm for the following problem:

Q-pattern:

$$\begin{aligned} \max_{\boldsymbol{\pi}} \quad & \sum_{k \in \mathcal{K}} U_k(\bar{R}_k) = \sum_{k \in \mathcal{K}} U_k\Big(\sum_{p \in \mathcal{P}} \phi_{kp} \pi_p \bar{r}_{kp} \Big) \\ \text{subject to} \quad & \sum_{p \in \mathcal{P}} \pi_p = 1, \end{aligned}$$

where $\phi_{kp} \in [0,1]$ is the probability that the user k is scheduled when pattern p is selected, i.e., $\phi_{kp} \cdot \pi_p = \bar{\pi}_{kp}$. For each pattern portion update epoch, i.e., every T_p slots, each BS n calculates the partial derivative $D_p^{(n)} \doteq \partial U^{(n)} / \partial \pi_p$ of per-cell utility $U^{(n)}$ with respect to the portion of pattern p and sends these information to the central coordinator.

$$D_p^{(n)} \doteq \frac{\partial U^{(n)}}{\partial \pi_p} = \sum_{k \in \mathcal{K}_n} U_k'(\bar{R}_k) \cdot \frac{\partial \bar{R}_k}{\partial \pi_p}, \tag{25}$$

where

$$\frac{\partial \bar{R}_k}{\partial \pi_p} = \phi_{kp} \bar{r}_{kp} = \frac{\bar{\pi}_{kp}}{\pi_p} \bar{r}_{kp}. \tag{26}$$

Note that three parameters ($\bar{R}_k$, $\bar{\pi}_{kp}$ and $\bar{r}_{kp}$) required to run this pattern portion update algorithm can be attained by the user scheduling algorithm. Then the central coordinator gathers information from all BSs and calculates the partial derivative of the network utility $D_p \doteq \partial U / \partial \pi_p$ by aggregating these partial derivatives of the local utility,

$$D_p \doteq \frac{\partial U}{\partial \pi_p} = \sum_{n \in \mathcal{N}} D_p^{(n)} \,, \quad p \in \mathcal{P}, \tag{27}$$

and updates the portion of reuse patterns following the ascent direction of network utility.

$$\boldsymbol{\pi} \leftarrow Proj_{\sum_{p \in \mathcal{P}} \pi_p = 1}, (\boldsymbol{\pi} + \gamma \mathbf{D}) \,. \tag{28}$$

Based on the updated portion of patterns, the central coordinator predetermines the sequence of patterns for next T_p slots that satisfies:

$$(\text{the total number of pattern } p) \;/\; T_p \approx \pi_p \,, \quad \forall p \in \mathcal{P}. \tag{29}$$

While there may be many strategies that leads to (29), a nice candidate is a random strategy. The central coordinator sequentially determines the sequence of patterns by rolling a P-dimensional die T_p times with probability of the pattern p being π_p. Once the sequence of patterns for next T_p slots is determined by the pattern portion change algorithm, then both BSs and users are informed of the sequence.

Now we develop the user scheduling algorithm under the fixed pattern given by the pattern portion change algorithm. From Lemma 3.4, for given pattern, the network-wide user scheduling problem can be decomposed into independent intra-cell user scheduling problems. Therefore, each BS needs to solve the following problem:

Q-scheduling:

$$\begin{aligned} \max_{(I_k(t), k \in \mathcal{K}_n)} \quad & \sum_{k \in \mathcal{K}_n} U_k'(\bar{R}_k(t-1)) I_k(t) r_{kp}(t) \\ \text{subject to} \quad & \sum_{k \in \mathcal{K}_n} I_k(t) \le 1. \end{aligned}$$

[3]For completeness, if a tie happens, the BS choose the lower indexed user.

TABLE I
COMPARISON BETWEEN JOINT OPTIMAL ALGORITHM (JOA) AND TIME-SCALE DECOMPOSED ALGORITHM (TDA)

	Joint optimal algorithm (JOA)	Time-scale decomposed algorithm (TDA)
Time-scale of algorithm	every slot	every slot (user scheduling) every T_p slot (pattern portion change)
Amount of feedback to each BS n at each slot	$\|\mathcal{K}_n\|\|\mathcal{P}_n\|$	$\|\mathcal{K}_n\|$
Amount of feedback to the central coordinator	$\sum_{n\in\mathcal{N}}\|\mathcal{K}_n\|\|\mathcal{P}_n\|$	$\sum_{n\in\mathcal{N}}\|\mathcal{P}_n\|$
Period of feedback to the central coordinator	1	T_p

The user scheduling algorithm solving **Q-scheduling** is straightforward. Each BS $n \in \mathcal{N}_p$ allowed to use the pattern p independently chooses the best user $k_n^*(t)$ among its associated user set $\mathcal{K}_n$, i.e., $I_{k_n^*}(t) = 1$:

$$k_n^*(t) = \arg\max_{k\in\mathcal{K}_n} U_k'(\bar{R}_k(t-1))r_{kp}(t), \quad \forall n \in \mathcal{N}_p, \quad (30)$$

and updates the following variables for the future purpose of the pattern portion change algorithm:

$$\begin{aligned}
\bar{R}_k(t) &= (1-\beta_1)\bar{R}_k(t-1) + \beta_1 I_k(t) r_{kp}(t)\ , \\
\bar{\pi}_{kp}(t) &= (1-\beta_2)\bar{\pi}_{kp}(t-1) + \beta_2 I_k(t)\ , \\
\bar{r}_{kp}(t) &= \begin{cases} (1-\beta_3)\bar{r}_{kp}(t-1) + \beta_3 r_{kp}(t), & \text{if } I_k(t)=1\ , \\ \bar{r}_{kp}(t-1), & \text{otherwise,} \end{cases}
\end{aligned}$$

where $\beta_1, \beta_2, \beta_3 > 0$ are small, averaging parameters; $\bar{R}_k(t)$, $\bar{\pi}_{kp}(t)$ and $\bar{r}_{kp}(t)$ are the average throughput of user k, the average fraction of time that user k is served with pattern p, and the average instantaneous data rate when the user k is served with pattern p, respectively.

Remark 4.1: For the user scheduling algorithm, it is required for each user to predict its potential data rate for the pattern in the next slot and report to its serving BS. We assume that each user can estimate the channel gains of different BSs from common pilot channels and send channel quality information to its serving BS through dedicate feedback channels. These functionalities are considered as a basic requirement in legacy and upcoming standard. Thus, each user can calculate its SINR for the pattern in the next slot by regarding the interference only from the activated BSs in the next slot as effective interference. Then, each user reports the predicted SINR (or corresponding potential data rate) to its serving BS. If the serving BS will not be activated in the next slot, the user does not need to send feedback at all.

Remark 4.2: There are two key differences between a recent algorithm [7] and the proposed algorithm. First, the referenced study additionally introduces a *virtual scheduler* to obtain the fraction of time in which the scheduler chooses user i for transmission in sub-band j (their notation: ϕ_{ij}). In the proposed algorithm, however, the fraction of time that user k is served with pattern p (our notation: $\bar{\pi}_{kp}$) is simply obtained using the *actual scheduler* without an extra algorithm. Second, the referenced study does not reflect the time-varying nature of the data rate available to user i in sub-band j (their notation: R_{ij}), as they assume that this rate does not change with time. In the proposed algorithm, the long-term average of the data rate of user k on pattern p (our notation: $\bar{r}_{kp}$) is not just the average of instantaneous data rate. We take the average of instantaneous data rate only if the user k is actually served by the scheduler. In other words, the value of $\bar{r}_{kp}$ in the proposed algorithm can reflect the multi-user diversity gain from exploiting the channel fluctuation.

C. Complexity Reduction and Its Price

The proposed time-scale decomposed algorithm still involves signalings from BSs to a central coordinator. However, it is possible to reduce feedback overhead significantly, as the periodicity of the feedback is *stretched* from every slot to every T_p slot. Moreover, the amount of feedback is reduced from $\left(\sum_{n\in\mathcal{N}}|\mathcal{K}_n||\mathcal{P}_n|\right)$ to $\left(\sum_{n\in\mathcal{N}}|\mathcal{P}_n|\right)$; i.e., it requires only BS-level feedback and not user-level channel feedback. The amount of feedback to each BS from its associated users at each slot is also reduced from $|\mathcal{K}_n||\mathcal{P}_n|$ and $|\mathcal{K}_n|$, as users need to send channel information only for a predetermined pattern. Table I compares the joint pattern selection and user scheduling algorithm with the proposed algorithms based on time-scale decomposition.

This complexity reduction for implementability comes at the cost of a performance gap with the joint optimal algorithm. This is due to the fact that the ICI management part in the decomposed algorithm cannot fully exploit instantaneous *inter-cell* channel variations; hence, only *intra-cell* channel variations are opportunistically utilized. Note that in the joint optimal algorithm, both pattern selection and user scheduling fully exploit both inter-cell and intra-cell time-varying channel conditions at a fast time scale.

As an example, consider a two cell network where two users are located at the edge of each cell. Their achievable rates are limited by severe ICI. The decomposed algorithm will find the following TDMA-like solution: BS 1 and BS 2 are exclusively active in order to mitigate the ICI, i.e., the portion of the pattern in which both BSs are active is nearly zero. However, suppose that both (time-varying) inter-cell channel gains from BS 1 (or 2) to the user in BS 2 (or 1) experience deep fading at some time slot. You can imagine this case as if there were a large wall between two cells. Subsequently the user in cell 1 (or 2) is not interfered by the BS transmission in cell 2 (or 1). Therefore, serving two users simultaneously is transiently optimal in this inter-cell deep fading case, whereas a pattern that only one BS is active is the solution for the average ICI mitigation case. The joint optimal algorithm can find this optimal solution by tracking this fast fading condition while the decomposed algorithm cannot. This is the cost for the complexity reduction, however, as shown in Section V, the performance gap becomes negligible in then absence of fast fading.

D. Construction method of the candidate pattern set $\mathcal{P}$

The number of all possible patterns in the network is an exponential function of the number of BSs, which may increase the complexity of the algorithm and reduce its practicality in the large network. However, most of these patterns are actually not used at all, thus it is required to select the essential candidate set of patterns out of all possible patterns.

Several papers [8], [15] have investigated on this topic. Bonald et al. [8] investigated an optimal transmit profile in the cellular network based on flow-level analysis. They concluded based on all their examples that the optimal capacity is attained by the use of two kinds of transmission patterns only: 1) one pattern is that all BSs are on and 2) the other patterns are that only the dominant interfering BS is switched off. Raman et al. [15] proposed a centralized spectrum server that finds an optimal schedule to maximize the average sum rate in general ad-hoc networks. They also conjectured that almost always only very few active transmission patterns are used as corroborated by their simulation results.

Encouraged by the observations in [8], [15], we make a practical guideline how to determine the candidate pattern set $\mathcal{P}$ as follows. The set should contain two kinds of mandatory patterns: 1) reuse-1 pattern: all BSs are active and 2) dominant patterns p: all neighboring BSs except the dominant interfering BS are activated by the pattern p. Any other appropriate patterns may be added in an optional manner if system designers want to increase the performance further, but the increment might be marginal.

V. Performance Evaluation

A. Simulation Setup

Two network configurations are considered for simulation-based performance evaluations: (i) a linear two-cell network and (ii) a two-tier multi-cell network with 19 cells. In both cases, the distance between BSs is set at 2km.

- *Linear two-cell network (see Fig. 1):* There are three patterns $\mathcal{P} = \{1, 2, 3\}$. Under pattern 1, both BSs are ON, and under pattern 2 (resp. 3), only BS 1 (resp. 2) is ON.
- *Two-tier multi-cell network (see Fig. 4):* 11 patterns (8 mandatory + 3 optional reuse-3 patterns) are considered. Under pattern 1, all BSs are ON. Under patterns 2∼4 (only one BS reuses the pattern among adjacent three BSs) or 5∼11 (six BSs reuse the pattern among adjacent seven BSs), a BS using these two types of patterns can expect ICI mitigation from the first-tier and from one of its neighboring cells, respectively.

In modeling the propagation environment, a path loss $-130 - 35\log_{10}(d_{km})$, log-normal shadowing with a standard deviation σ_s=8dB and Jakes' Rayleigh fading (3km/h) for fast fading are adopted. In subsection V-B, we evaluate the performance of the proposed algorithm with and without fast fading cases. In the case with fast fading, the channel varies over the time since all the above mentioned models are considered. On the other hand, in the case without fast fading, we consider only the path loss and shadowing models except Jakes' fading. Therefore, the channels remain stable during the simulation time. The channel bandwidth and the time-slot length are set at 10MHz and 5ms, respectively. The maximum transmission powers of the BSs are all identical to 20W. The other parameters for the simulations follow the suggestions in the IEEE 802.16m evaluation methodology document [16]. All users have a logarithmic utility function, i.e., (w, α) = (1,1). The pattern update period T_p is set at 500 and the step size is chosen to be a typically small value, i.e., $\beta_1 = \beta_2 = \beta_3 = \gamma = 0.001$. We have tested other values of T_p, β_1, β_2, and β_3, and the similar results were obtained. Simulations were run of over 50000 slots.

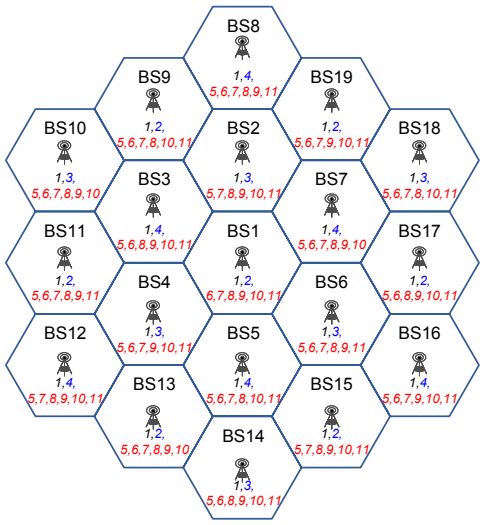


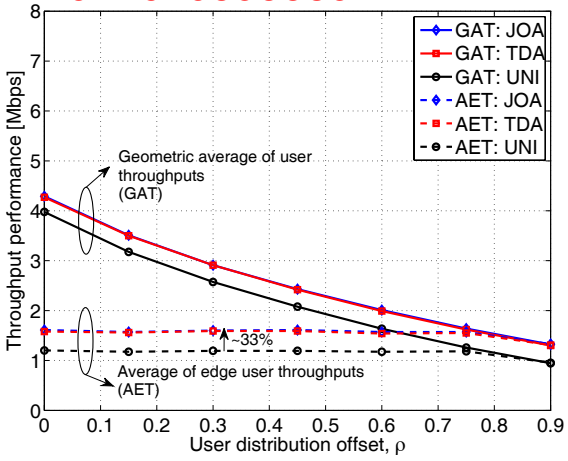
Fig. 4. Two-tier multi-cell network composed of 19 cells.

To evaluate the performance under various user distribution scenarios, we introduce a variable, so-called, "user distribution offset" $\rho \in [0, 1]$. It adjusts the minimum distance between the BS and the user to $\rho \times$ (cell radius). Users in each cell are randomly distributed with this minimum distance restriction. For example, if $\rho = 0$, users are uniformly generated over the entire cell. On the other hand, if ρ becomes 1, users swarm the edge areas of cells.

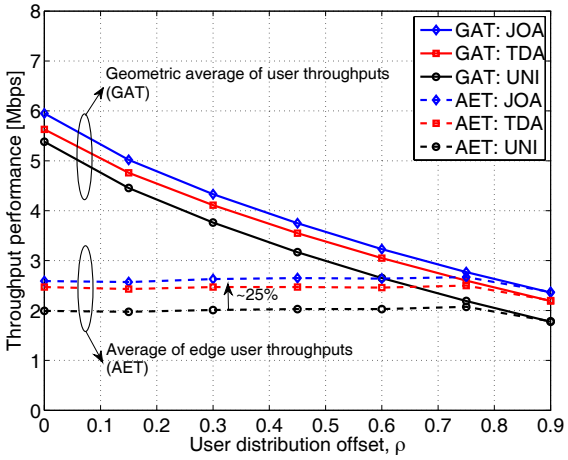
Simulation results of the following three algorithms are presented:(i) a *conventional universal reuse scheme* (UNI), in which all BSs in the network are always active without any ICI management, (ii) *the joint optimal algorithm* (JOA) in (19), and (iii) *the algorithm based on time-scale decomposition* (TDA) in (20)∼(24). As performance metrics, the geometric average of user throughputs (GAT) and the average of edge user throughputs (AET) are used. We use GAT because maximizing this metric is equivalent to the system objective (sum of log throughputs). AET is a measure of cell edge performance that is defined as the average throughput of the users located at the cell edges. In simulation, we treat "edge users" as those who are more than 800m away from their serving BSs, otherwise termed "center users".

B. Two-Tier 19-Cell Network Case

We first consider the two-tier multi-cell network with 19 cells, where each cell has ten users. Without fast fading, as shown in Fig. 5(a), both JOA and TDA perform similarly. Compared to UNI, GAT and AET of JOA and TDA increases by 10∼33% (depending on the user distribution) and by



(a) Without fast fading



(b) With fast fading

Fig. 5. Throughput performances of three algorithms: joint optimal algorithm (JOA), time-scale decomposed algorithm (TDA) and universal reuse (UNI).

33%. A higher performance gain was observed when the user distribution offset is larger (i.e., more users are located at the cell edges), which is due to the fact that ICI management mainly targets for performance improvement of edge users. With fast fading, as shown in Fig. 5(b) however, as discussed in subsection IV-C, a performance gap between JOA and TDA exists due to the loss in opportunism in TDA. However, TDA still outperforms UNI in terms of both GAT (5∼25% depending on user distribution) and AET (25%). It is noteworthy that TDA can attain more than 1/2 (at $\rho = 0$) and up to 2/3 (at $\rho = 0.9$) of the GAT performance gain that can be achieved by JOA.

Fig. 6 shows the convergence of 11 pattern portions when offset ρ is equal to 0.3. As you can see, the pattern portions converge quickly in 10∼15 iterations. Even in different network configurations (with a different number of BSs and a different values of offset), we could see similar convergence trends, e.g., typically within 10∼20 iterations.

In our simulation, the proportional fairness ($\alpha = 1$) is considered as our system objective. Although additional simulation results with other values of α are not included in this paper due to space limitations, higher performance gains were obtained for higher values of α (i.e., higher fairness objectives), as increasing the throughput of edge users is essential in order to achieve higher fairness. Thus, ICI management can offer conspicuous improvement in a network pursuing fairness-oriented system objectives.

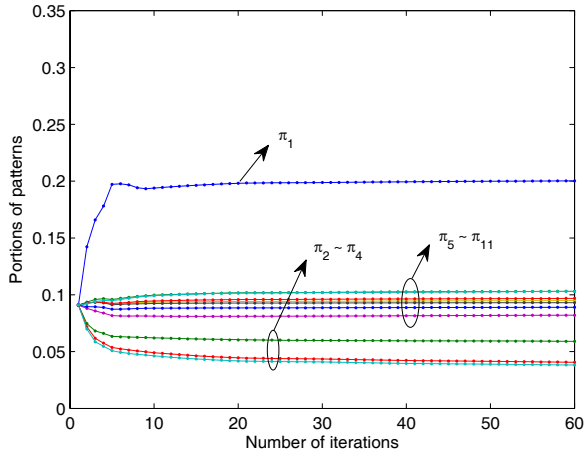


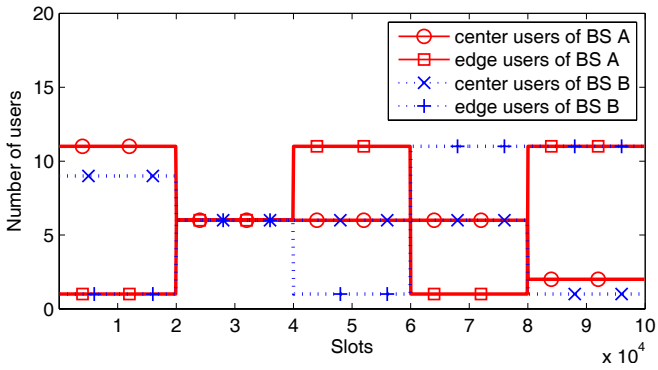
Fig. 6. Rate of convergence.

C. Adaptation Test

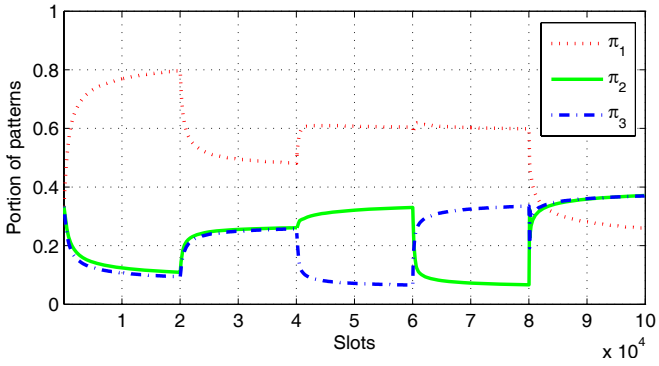
To test adaptation to dynamic changes in user loads, the linear two-cell network was used. The user load distribution was varied every 2×10^4 slots (=100 sec) as shown in Fig. 7(a). The TDA was compared with other three *static* schemes: RF1, RF2 and MIXED. RF1 and RF2 correspond to those with reuse factors 1 and 2, respectively. RF1 and RF2 are incorporated in the MIXED scheme, which operates as a RF1 during the half portion of the time and RF2 during the other half portion of the time. The RF1, RF2 and MIXED schemes can be regarded as those with static pattern portions (π_1, π_2, π_3) = (1, 0, 0), (0, 0.5, 0.5) and (0.5, 0.25, 0.25), respectively.

Fig. 7(b) shows the pattern portion adaptation characteristics of TDA to a dynamic user load distribution. In the period 1 ($0 \sim 2 \times 10^4$ slots), as there are relatively many users around the cell center, the portion of pattern 1 increases up to 80%, which is nearly identical to that of RF1. On the other hand, in the another ($8 \sim 10 \times 10^4$ slots), as there are relatively many users around the cell edge, the portion of pattern 1 decreases up to 20%, which is nearly identical to that of RF2. Users are equally divided into center and edge regions in the period 2 ($2 \sim 4 \times 10^4$ slots) so that the pattern portion becomes (0.5, 0.25, 0.25) like as MIXED. In periods 3 and 4, there is the same number of center users in each cell, but the number of edge users are different; i.e., heterogeneous load distribution. In period 3 (resp. 4), BS A (resp. B) exceeds the number of edge users. As expected, π_2 (resp. π_3) increases while π_3 (resp. π_2) decreases.

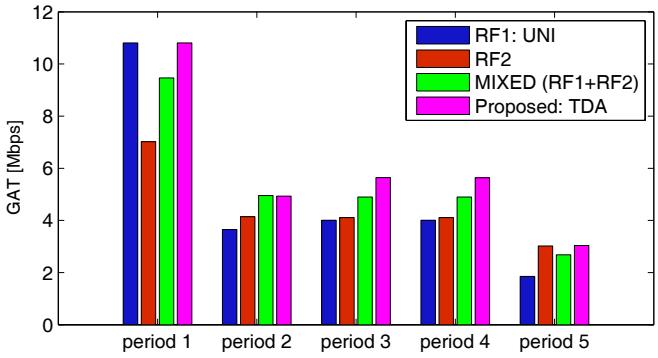
Fig. 7(c) shows the GAT performance comparison period by period. Static schemes perform well only if the user load distribution fits their patterns, for example, RF1 in period 1, RF2 in period 5 and MIXED in period 2. In the cases in-



(a) Dynamic change of user load distribution



(b) Pattern portion adaption to the dynamic user load distribution



(c) GAT performance comparison period by period

Fig. 7. Adaptiveness to dynamic user load.

volving other distributions, however, performance degradation is inevitable. Compared to these static schemes, the proposed TDA scheme can adapt to the dynamic user load distribution and find the optimal portion of patterns. Therefore, it always achieves the best performance in all cases.

D. Imbalance Load Scenario: Association Control vs. Interference Control

We also test the performance in the linear two-cell network for imbalanced loads. Two users were located 900m away from BS 1 and $(2 \times LI)$ users were located 900m away from BS 2, respectively, where LI quantifies the load imbalance. Under

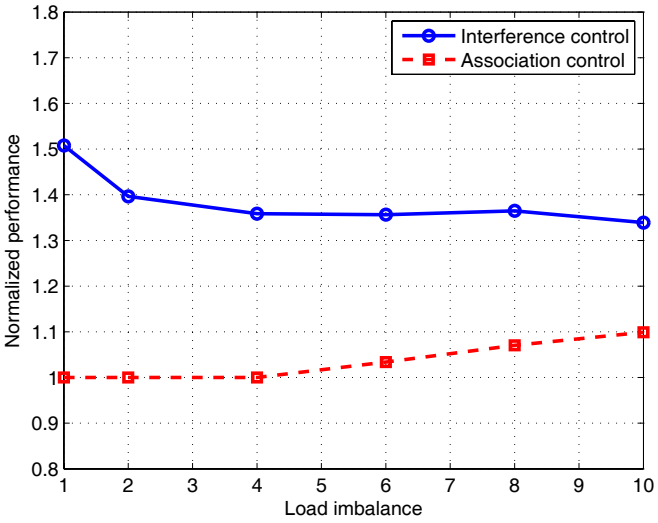


Fig. 8. Association control vs. interference control under imbalanced load.

this load imbalance scenario, the performance of the following two different approaches are compared: (i) *Association control: a load-aware handover* in [9] and (ii) *Interference control*: the proposed TDA. As a baseline, we considered the GAT performance of UNI and normalized the performance of the two approaches above by UNI.

Originally, users are associated with the closest BS offering the best signal strength. In the case of the association control approach, however, if the expected throughput measure in [9] from the other BS is greater than that from the current BS, then the user changes its association. When the LI value is small, users do not change their associations. When LI is increased to more than 6, association changes from the hot-spot cell (BS 2) and the under-loaded cell (BS1) occur (moving one, two and three users at LI=6, 8 and 10, respectively) according to the load-aware handover in [9]. As shown in Fig. 8, however, the gain from the association control is marginal.

On the other hand, using the interference control approach, we can implicitly resolve the load imbalance by preventing the hot-spot cell (BS 2) from being turned off, i.e., provide more interference-free transmission opportunities compared to BS 1. In brief, the interference control approach originally developed for ICI mitigation can also resolve the load imbalance as well, and the improvement of interference control is superior to that of the association control.

VI. Conclusion

In this paper, we have focused on the problem of joint ICI management and user scheduling in multi-cell networks. It was shown that the joint optimal algorithm is too complex in terms of computational and signaling overhead to be implemented in practical systems. To overcome this complexity and make the algorithm practical, we decomposed the original optimization problem into two sub-problems, where ICI management is run at a slower time scale compared to user scheduling. This time-scale decomposition stems from a design rationale in which ICI management may not have to track fast changing dynamics, and it may suffice to attain much gain simply by running it based only on macroscopic network changes. We

empirically show that even with such a slow tracking of the system dynamics at the ICI management, our algorithm can achieve high performance gain compared to the conventional universal reuse scheme. Moreover, it is practically implementable compared to the very complex optimal algorithm.

APPENDIX

A. Proof of Lemma 3.1

Let λ_0 and λ_{kp} denote the Lagrangian multiplier associated with the constraint (8) and constraint (9), respectively. Then the Lagrangian function is given by:

$$\mathcal{L}(\pi,\lambda) = \sum_{k\in\mathcal{K}_1} U_k(\bar{R}_k) - \lambda_0\Big(\sum_{p\in\mathcal{P}_1}\sum_{k\in\mathcal{K}_1}\frac{\pi_{kp}}{\chi_p} - 1\Big) + \sum_{k\in\mathcal{K}_1}\sum_{p\in\mathcal{P}_1}\lambda_{kp}\pi_{kp}.$$

The problem **Q-symmetric** is a convex optimization so the necessary and sufficient conditions for optimality are given by Karush-Kuhn-Tucker (KKT) conditions [17]:

i) $\lambda_0 \geq 0$ and $\lambda_{kp} \geq 0,\ \forall k \in \mathcal{K}_1, \forall p,$ (31)

ii) $\lambda_0\Big(\sum_{p\in\mathcal{P}_1}\sum_{k\in\mathcal{K}_1}\frac{\pi_{kp}}{\chi_p} - 1\Big) = 0$ and $\lambda_{kp}\pi_{kp} = 0,\ \forall k, \forall p,$ (32)

iii) $\frac{\partial \mathcal{L}}{\partial \pi_{kp}} = U_k'(\bar{R}_k) r_{kp} - \frac{\lambda_0}{\chi_p} + \lambda_{kp} = 0,\ \forall k \in \mathcal{K}_1, \forall p.$ (33)

Substituting (33) into (32) and (31) yields the following conditions, (34) and (35), respectively.

$$\Big[\frac{\lambda_0}{\chi_p} - U_k'(\bar{R}_k) r_{kp}\Big]\pi_{kp} = 0, \quad \forall k \in \mathcal{K}_1, \forall p \in \mathcal{P}_1, \quad (34)$$

$$\lambda_0 \geq U_k'(\bar{R}_k)\chi_p r_{kp}, \quad \forall k \in \mathcal{K}_1, \forall p \in \mathcal{P}_1. \quad (35)$$

Let $p^*(k) = \arg\max_p \chi_p r_{kp}$ denote the most efficient pattern of user k. By (35), there are two cases.

Case 1: If $\lambda_0 > U_k'(\bar{R}_k)\chi_{p^*(k)} r_{kp^*(k)}$, then $\Big[\frac{\lambda_0}{\chi_p} - U_k'(\bar{R}_k) r_{kp}\Big] > U_k'(\bar{R}_k) r_{kp}\Big(\frac{\chi_{p^*(k)} r_{kp^*(k)}}{\chi_p r_{kp}} - 1\Big) > 0$. Consequently, (34) holds only if $\pi_{kp} = 0, \forall k, \forall p$, however, this is not a reasonable solution.

Case 2: If $\lambda_0 = U_k'(\bar{R}_k)\chi_{p^*(k)} r_{kp^*(k)}$, then $\Big[\frac{\lambda_0}{\chi_p} - U_k'(\bar{R}_k) r_{kp}\Big] = U_k'(\bar{R}_k) r_{kp}\Big(\frac{\chi_{p^*(k)} r_{kp^*(k)}}{\chi_p r_{kp}} - 1\Big)$. Consequently, (34) holds only if $\pi_{kp} \geq 0$ if $p = p^*(k)$, and 0 otherwise. This completes the proof of Lemma 3.1. ■

B. Proof of Lemma 3.2

By the Lemma 3.1, we can rewrite the condition (33) for $p = p^*(k)$ as follows:

$$U_k'(\bar{R}_k) r_{kp^*(k)} - \frac{\lambda_0}{\chi_{p^*(k)}} = 0. \quad (36)$$

The derivative for the generalized (w,α)-fair utility is given by $U_k'(\bar{R}_k) = w_k/\bar{R}_k^\alpha = w_k/(\pi_{kp^*(k)} r_{kp^*(k)})^\alpha$. Substituting this derivative into (36) yields the following optimal fraction of time for user-patterns:

$$\pi_{kp^*(k)} = \big(w_k \chi_{p^*(k)} r_{kp^*(k)}^{1-\alpha}/\lambda_0\big)^{1/\alpha}. \quad (37)$$

Since $\pi_{kp} = 0$ for $p \neq p^*(k)$, we can rewrite the condition (32) as follows:

$$\sum_{p\in\mathcal{P}_1}\sum_{k\in\mathcal{K}_1}\frac{\pi_{kp}}{\chi_p} - 1 = \sum_{p\in\mathcal{P}_1}\sum_{k\in\mathcal{K}_{1p}}\frac{\pi_{kp}}{\chi_p} - 1 = 0, \quad (38)$$

where $\mathcal{K}_{1p}$ is the set of users whose most effective pattern having the highest effective rate is p, i.e., $p = p^*(k) = \arg\max_p \chi_p r_{kp}$ if $k \in \mathcal{K}_{1p}$. By plugging (37) into (38), we can have the optimal value of λ_0:

$$\lambda_0 = \Big(\sum_{p\in\mathcal{P}_1}\sum_{k\in\mathcal{K}_{1p}} w_k^{1/\alpha}\chi_{p^*(k)}^{\frac{1-\alpha}{\alpha}} r_{kp^*(k)}^{\frac{1-\alpha}{\alpha}}\Big)^\alpha. \quad (39)$$

This completes the proof of Lemma 3.2. ■

C. Proof of Lemma 3.4

For the given pattern p, i.e., $X_p(t) = 1$, we can rewrite (15) as follows,

$$\begin{aligned}\Delta U(t) &= \sum_{n\in\mathcal{N}}\sum_{k\in\mathcal{K}_n} U_k'(\bar{R}_k(t-1)) I_k(t) r_{kp}(t) \\ &= \sum_{n\in\mathcal{N}_p}\sum_{k\in\mathcal{K}_n} U_k'(\bar{R}_k(t-1)) I_k(t) r_{kp}(t) \\ &= \sum_{n\in\mathcal{N}_p}\Delta U_n(t),\end{aligned}$$

where the second equality holds from (2), $I_k(t) = 0,\ n \notin \mathcal{N}_p$ and $\Delta U_n(t) = \sum_{k\in\mathcal{K}_n} U_k'(\bar{R}_k(t-1)) I_k(t) r_{kp}(t)$. As $U_k'(\bar{R}_k(t-1))$ and $r_{kp}(t)$ are given parameters at slot t, we only have to investigate decencies among $I_k(t)$. Since the constraints (2) on $I_k(t)$ do not play a role across different BSs, $\Delta U_n(t)$ are mutually independent. Therefore, solving the original problem is equivalent to maximize $\Delta U_n(t)$ independently for each BS $n \in \mathcal{N}_p$.

$$\max_{\mathbf{I}(t)} \quad \Delta U_n(t) = \sum_{k\in\mathcal{K}_n} U_k'(\bar{R}_k(t-1)) I_k(t) r_{kp}(t)$$
$$\text{subject to} \quad \sum_{k\in\mathcal{K}_n} I_k(t) \leq 1,$$

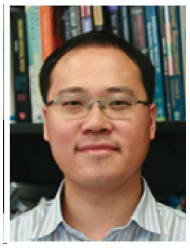
This problem can be further simplified as (18), which completes the proof of Lemma 3.4. ■

REFERENCES

[1] WiMAX Forum, "Mobile WiMAX—part I: a technical overview and performance evaluation," Aug. 2006.
[2] A. Gjendemsj, D. Gesbert, G. E. Oien, and S. G. Kiani, "Binary power control for sum rate maximization over multiple interfering links," *IEEE Trans. Wireless Commun.*, vol. 7, no. 8, pp. 3164–3173, Aug. 2008.
[3] S. Das, H. Viswanathan, and G. Rittenhouse, "Dynamic load balancing through coordinated scheduling in packet data systems," in *Proc. IEEE INFOCOM*, San Francisco, CA, Mar. 2003.
[4] J. Cho, J. Mo, and S. Chong, "Joint network-wide opportunistic scheduling and power control in multi-cell networks," in *Proc. IEEE WoWMoM*, San Francisco, CA, June 2007.
[5] 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, July 2009.
[6] B. Rengarajan and G. de Veciana, "Network architecture and abstractions for environment and traffic aware system-level coordination of wireless networks: the downlink case," in *Proc. IEEE INFOCOM*, Phoeniz, AZ, Apr. 2008.
[7] A. L. Stolyar and H. Viswanathan, "Self-organizing dynamic fractional frequency reuse for best-effort traffic through distributed inter-cell coordination," in *Proc. IEEE INFOCOM*, Rio de Janeiro, Brazil, Apr. 2009.

[8] T. Bonald, S. Borst, and A. Proutière, "Inter-cell scheduling in wireless data networks," in *Proc. European Wireless*, Cyprus, Greece, Apr. 2005.
[9] A. Sang, X. Wang, M. Madihian, and R. D. Gitlin, "Coordinated load balancing, handoff/cell-site selection, and scheduling in multi-cell packet data systems," in *Proc. ACM MobiCom*, Philadelphia, PA, Sep. 2004, pp. 302–314.
[10] T. Bu, L. Li, and R. Ramjee, "Generalized proportional fair scheduling in third generation wireless data networks," in *Proc. IEEE INFOCOM*, Barcelona, Spain, Apr. 2006.
[11] F. Kelly, A. Maullo, and D. Tan, "Rate control in communication networks: shadow prices, proportional fairness and stability," *J. Operational Research Society*, vol. 49, pp. 237–252, July 1998.
[12] 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.
[13] A. L. Stolyar, "On the asymptotic optimality of the gradient scheduling algorithm for multiuser throughput allocation," *Operations Research*, vol. 53, no. 1, pp. 12–25, Jan. 2005.
[14] H. J. Kushner and P. A. Whiting, "Convergence of proportional-fair sharing algorithms under general conditions," *IEEE Trans. Wireless Commun.*, vol. 3, no. 4, pp. 1250–1259, July 2004.
[15] C. Raman, R. D. Yates, and N. B. Mandayam, "Scheduling variable rate links via a spectrum server," in *Proc. IEEE DySPAN*, Baltimore, MD, Nov. 2005.
[16] IEEE 802.16m-08/004r5, "IEEE 802.16m Evaluation Methodology Document (EMD)," Jan. 2009.
[17] S. Boyd and L. Vandenberghe, *Convex Optimization*, 1st edition. Cambirdge University Press, 2004.

Kyuho Son (S'03-M'10) received his B.S., M.S. and Ph.D. degrees all in the Department of Electrical Engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2002, 2004 and 2010, respectively. He is currently a post-doctoral research associate in the Department of Electrical Engineering at the University of Southern California. His current research interests include interference management in heterogeneous cellular networks, green wireless networking and network economics. He served as the Web Chair of the 7th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt 2009).



Yung Yi (S'04-M'06) received his B.S. and the M.S. in the School of Computer Science and Engineering from Seoul National University, South Korea in 1997 and 1999, respectively, and his Ph.D. in the Department of Electrical and Computer Engineering at the University of Texas at Austin in 2006. From 2006 to 2008, he was a post-doctoral research associate in the Department of Electrical Engineering at Princeton University. Now, he is an assistant professor at the Department of Electrical Engineering at KAIST, South Korea. His current research interests include the design and analysis of computer networking and wireless communication systems, especially congestion control, scheduling, and interference management, with applications in wireless ad hoc networks, broadband access networks, economic aspects of communication networks (aka network economics), and green networking systems. He has been serving as a TPC member at various conferences such as ACM Mobihoc, Wicon, WiOpt, IEEE Infocom, ICC, Globecom, ACM CFI, ITC, the local arrangement chair of WiOpt 2009 and CFI 2010, and the networking area track chair of TENCON 2010.

Song Chong (S'93-M'95) received the B.S. and M.S. degrees in Control and Instrumentation Engineering from Seoul National University, Seoul, Korea, in 1988 and 1990, respectively, and the Ph.D. degree in Electrical and Computer Engineering from the University of Texas at Austin in 1995. Since March 2000, he has been with the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, where he is a Professor and the Head of the Communications and Computing Group of the department. Prior to joining KAIST, he was with the Performance Analysis Department, AT&T Bell Laboratories, New Jersey, as a Member of Technical Staff. His current research interests include wireless networks, future Internet, and human mobility characterization and its applications to mobile networking. He has published more than 90 papers in international journals and conferences.

He is an Editor of *Computer Communications* journal and *Journal of Communications and Networks*. He has served on the Technical Program Committee of a number of leading international conferences including IEEE INFOCOM and ACM CoNEXT. He serves on the Steering Committee of WiOpt and was the General Chair of WiOpt '09. He is currently the Chair of Wireless Working Group of the Future Internet Forum of Korea and the Vice President of the Information and Communication Society of Korea.