Optimal Joint User Association and Resource Allocation in Heterogeneous Networks via Sparsity Pursuit

Quan Kuang, Wolfgang Utschick, and Andreas Dotzler

Abstract

This paper studies the joint user association and resource allocation in heterogeneous networks (HetNets) from a novel perspective, motivated by and generating the idea of fractional frequency reuse. By treating the multi-cell multi-user resource allocation as resource partitioning among multiple reuse patterns, we propose a unified framework to analyze and compare a wide range of user association and resource allocation strategies for HetNets, and provide a benchmark of ultimate limit on network performance. The enabling mechanisms are a novel formulation to consider all possible patterns or any pre-defined subset of patterns and an efficient sparsity-pursuit algorithm. More importantly, in view of the fact that multi-cell resource allocation is very computational demanding, our framework provides a systematic way to trade off performance for the reduction of computational complexity by restricting the candidate patterns to a small number of feature patterns. Relying on the sparsity-pursuit capability of the proposed algorithm, we develop practical guideline to identify the feature patterns in the given HetNet. Our treatment is very general in that it covers the case where users are allowed to associate with multiple base stations and the more restrictive situation where the number of serving base stations for each user is limited to one or a given number. Numerical results show that the identified feature patterns can significantly improve the existing strategies, and jointly optimizing the user association and resource allocation indeed brings considerable gain.

Index Terms

heterogeneous networks, user association, inter-cell interference coordination, interference management, fractional frequency reuse, reuse pattern, frequency allocation, channel assignment, sparsity pursuit,
Frank-Wolfe algorithm, conditional gradient method, 3GPP LTE, range expansion, eICIC, almost blank subframe (ABS), resource allocation

I. INTRODUCTION

The heterogeneous network (HetNet), where low-power low-complexity base-stations (BSs) are overlaid with conventional macro BSs, is being considered as a promising paradigm for increasing system capacity and coverage in a cost-effective way. Due to Orthogonal Frequency Division Multiple Access (OFDMA) mechanism adopted in these networks, intra-cell interference is nearly null. However, the inter-cell interference (ICI) potentially introduced by hierarchical layering of cells becomes a fundamental limiting factor to the HetNet performance. One way to control the ICI is to allocate the time/frequency resources in an intelligent manner across multiple cells, determining which BSs should transmit on which channels, and at which time instance.

This resource allocation problem in HetNets is highly coupled with the user traffic distribution and user association policy. In a macro-only cellular network, the user association can be decoupled from the resource allocation and the user is simply associated to the BS with strongest downlink signal strength. In a HetNet, however, this association policy will lead to the case where the macro cells become resource constrained while the small cells are extremely underutilized, due to the large discrepancy in their transmit power. To balance the load, a user may be connected to a small cell even though the received power from a macro BS is higher. However, this may cause severe interference and overwhelm the cell-splitting gain eventually, if the radio resources are not carefully partitioned among cells. Clearly, the resource allocation and user association should be optimized jointly [1].

This paper studies the joint user association and multi-cell resource allocation from a novel perspective, motivated by and generalizing the idea of fractional frequency reuse (FFR) [2], [3]. The basic mechanism of FFR is to pre-define a set of reuse patterns. Each of these patterns determines a particular combination of ON/OFF activities of all BSs. For example, the reuse-1 pattern simply activates all the BSs, and a reuse-3 pattern activates one BS among neighboring three BSs. In conventional FFR methods [3], [4], we can allocate a certain portion of system spectrum to reuse-1 pattern (which is referred to as universally reused subbands), and then divide the rest of spectrum evenly among three reuse-3 modes. As a result, each cell can schedule less vulnerable users (e.g., at cell-center) to the universal subbands and cell-edge users to the subbands with lower number of active BSs. Our framework generalizes this idea by developing mechanisms to consider any pre-defined or all possible patterns. Unlike the majority of related works that deal with multi-cell channel assignment directly (see for example, [5], [6] and references...
Fig. 1. An illustration of the multi-pattern frequency allocation in a HetNet consisting of one macro (Cell 1) and two small cells (Cells 2 and 3). To represent a particular pattern, we use 0 at position $i$ of a row vector to indicate Cell $i$ is OFF, and 1 if Cell $i$ is ON. Bar chart (a) gives a result of frequency allocation among all 7 patterns, where $\{\pi_1, \cdots, \pi_7\}$ are the allocation profile. Result of (a) can be directly translated into frequency allocation among different cells as shown in (b), (c), and (d) respectively, where forbidden frequency channels are indicated in white color.

therein), we treat the time/frequency allocation among different cells as time/frequency partitioning among multiple patterns. Fig. 1 illustrates how the results of allocating frequency resources among patterns can be translated directly into multi-cell frequency allocation, where a network consisting of 3 cells and $2^3 - 1 = 7$ possible patterns are considered.

The advantage of our treatment is that it enables us to develop a unified framework to study various user association and resource allocation strategies for HetNets. Using our framework, a wide range of existing strategies can be analyzed and compared in a unified way by restricting the candidate patterns to a certain set of pre-selected ones. Our framework is also able to investigate the strategies in which user association is either jointly optimized with resource allocation, or decoupled from the resource allocation and performed by some simple rules.

More importantly, as multi-cell resource allocation is very computationally demanding due to the very large number of variables, it is desirable to reduce the computational effort at the expense of sacrificing the performance as little as possible. Our framework provides a systematic way to trade off performance for the reduction of computational complexity. This can be done by restricting the number of candidate patterns. However, unlike in the conventional homogeneous network where the mixture of reuse-1 and
reuse-3 patterns is a natural choice, candidate pattern selection in a HetNet is no longer a simple task, due to the irregular cell location and overlaid cell deployment. Our framework is able to identify the feature patterns for the given network.

We build our framework on the sparsity pursuit optimization. Specifically, we first formulate the unrestricted user association and multi-pattern resource allocation problem. The formulation is general enough to consider all possible patterns and users are allowed to associate with multiple BSs. Although the problem is convex, solving it is highly nontrivial due to the extremely high problem dimension. We then develop an efficient sparsity-pursuit algorithm (Algorithm 1) that is able to find optimal sparse solutions. Relying on the sparsity-pursuit capability of Algorithm 1, we develop practical guideline to identify the feature patterns in the given HetNet. We further propose an algorithm (Algorithm 2) to take into account the practical situations where the number of serving BSs for each user is restricted to one (i.e., single-BS association) or a given number. Finally, we compare various user association and resource allocation strategies using our framework. We show that the identified feature patterns can significantly improve the existing strategies, and jointly optimizing the user association and resource allocation indeed brings considerable gain.

A. Related work

FFR has attracted lots of research efforts from both academia [7]–[10] and industrial standardizations [11]. These frequency reuse schemes have been implemented in a static manner during the network planning phase [3], [8], [9], or adaptively according to the time-variations in cell traffic loads [7], [10]. The joint power and frequency allocation for OFDMA FFR has also been studied, for example in [6] and references cited therein.

On the other hand, there has been an increasing interest in user association problems. While the earlier works and some recent ones focus on the minimization of the total transmit power [12]–[15], the current trends are to optimize user association for load balancing in HetNets as described in [16], [17]. Meanwhile, simple “range expansion” techniques have been introduced in Third Generation Partnership Project (3GPP) standardization bodies to off-load macro users to small cells by adding a positive bias to the downlink signal strength of small BSs during the cell selection [18]. The off-loading effect of range expansion techniques has been extensively studied, for example, in [18]–[20].

However, despite their coupled nature, the joint user association and frequency allocation for FFR has been less explored. In [21], the authors studied a dynamic user association problem in a FFR network. However, the frequency partitioning is assumed given beforehand and fixed. By assuming that all cells use
all spectrum (reuse-1 pattern), [16] investigated a joint user association and \textit{intra-cell} resource allocation problem. Our formulation generalizes their formulation by taking into account the \textit{inter-cell} resource allocation as well. In particular, we study multi-pattern resource allocation as an effective mean of reducing the inter-cell interference, together with the user association strategy and intra-cell resource allocation.

In [1], the joint multi-cell channel allocation and user association was studied. However, their study was restricted to three pre-defined resource allocation strategies, namely, orthogonal deployment, co-channel deployment, and partially shared deployment. Our framework generalizes their studies in that it allows us to consider any pre-defined or all possible strategies (i.e., reuse patterns) when the user association is also optimized. We provide a unified framework to analyze and compare the performance of various resource allocation strategies, including but not limited to, those three defined in [1]. More importantly, our formulation provides us a vehicle to identify the essential strategies (reuse patterns) that give better performance than those defined in [1] by intuitions.

The joint user association and interference coordination via time-domain almost blank subframes (ABSs) was investigated in [23], [24], where macro BSs do not transmit any data over certain subframes periodically to configure ABSs such that pico cells can schedule vulnerable users on those resources with reduced interference. Each macro BS is assumed to have the same blank subframes in [23], [24]. As will become clear next, this strategy can be easily analyzed using our framework by defining only two patterns: all-ON and only-pico-ON. Moreover, our framework covers more general case where each macro BS may have different blank subframes.

There are some other works related to joint user association and interference management. For example, in [22], the authors jointly optimized the user association, power and frequency allocation. However, the formulation and methodology only apply to the uplink sum rate maximization. Note that reuse pattern selection has been studied in a \textit{Time Division Multiple Access (TDMA)} macro-only network in [25], which is different from our formulation. Besides, the HetNet deployment requires new criterion to select essential patterns.

In our preliminary work [26], a joint user association and reuse pattern selection problem was formulated where only single-BS association is allowed and each BS serves multiple users by a round-robin scheduler, allocating all its spectrum to a single user at a given time. Since the resulting problem is nonconvex and combinatorial, a heuristic approach based on Tabu search was adopted to solve the problem. However, it is not clear how far the solution is from the optimum. In this paper, we obtain a convex formulation for more general assumptions and global solution is found to such problem.
B. Outline of the paper

Section II introduces the system model. The unrestricted user association and pattern selection problem is formulated Section III. Section III also presents the properties of the formulated problem and the details of the proposed sparsity-pursuit algorithm for solving the problem. In Section IV, we consider the additional cardinality constraints for user association, and develop another algorithm to solve such problem. In Section V, we study the feature pattern identification using our sparsity-pursuit algorithm, and compare various strategies using our framework.

II. SYSTEM MODEL

We consider a downlink OFDMA HetNet, where a number of small cells are embedded in the conventional macro cellular network. The set of all cells is denoted as $\mathcal{B}$, and the cardinality $|\mathcal{B}| = B$. A set of users $\mathcal{K}$, with $|\mathcal{K}| = K$, are distributed in the network.

We use $g_{bk,n} \triangleq \sqrt{G_{bk}h_{bk,n}}$ to denote the channel gain between BS $b$ and user $k$ at subcarrier $n$, where $G_{bk}$ is the large-scale channel gain including path loss and shadowing, $h_{bk,n}$ accounts for the small-scale fading. We assume $\{h_{bk,n}, \forall b, \forall k, \forall n\}$ are independent and identically distributed (i.i.d.). Hence, the ergodic rate of user $k$, if it is exclusively served by $b$-th BS under pattern $i$, can be written as

$$r_{kbi} = W \mathbb{E}_{h_{k,n}} \left[ \log_2 \left( 1 + \frac{P_i^b G_{bk}||h_{bk,n}||^2}{\sigma^2 + \sum_{l \neq b} P_i^l G_{lk}||h_{lk,n}||^2} \right) \right] \text{ bit/s}$$

where $W$ is the system bandwidth, $h_{k,n} \triangleq (h_{1k,n}, h_{2k,n}, \ldots, h_{Bk,n})$ with $h_{bk,n}$ denoting the fast fading coefficient between BS $b$ and user $k$ over an arbitrary frequency subcarrier, $P_i^b$ and $\sigma^2$ are the transmit power of BS $b$ per Hz and the received noise power per Hz, respectively. We assume if BS $b$ is active transmit power is evenly distributed across all frequencies, i.e., $P_i^b = P_{tx}^b / W$ where $P_{tx}^b$ is the maximum transmit power of BS $b$; otherwise $P_i^b = 0$ if BS $b$ is muted under $i$-th pattern. We further assume uniformly distributed noise effect over frequencies. Hence the ergodic rate of (1) is uniform across the whole bandwidth for any pattern.

Given the set of all possible reuse patterns $\mathcal{I}$, $|\mathcal{I}| = |\mathcal{I}|$, $\{r_{kbi}\}$ can be pre-calculated using (1) and treated as constants. The resource allocation strategy is to distribute available orthogonal resources among different patterns. Note that this can be done either in the frequency domain in terms of bandwidth, or

1 Cell and BS are used interchangeably in this paper.

2 This implies that BS $b$ may not transmit at its maximum power $P_{tx}^b$. The actual transmit power is $P_i^b \times W_b$, where $W_b$ denotes the bandwidth allocated to BS $b$ after multi-cell frequency allocation. Thus, the power control is implicitly considered in our model under the assumption of uniform power distribution.
in the time domain in terms of subframes. Let \( \pi \triangleq (\pi_1, \ldots, \pi_i, \ldots, \pi_I) \in \Pi \) be the allocation profile, where \( \pi_i \) represents the fraction allocated to pattern \( i \) and \( \Pi = \{ \pi : \sum_i \pi_i = 1, \pi_i \geq 0, \forall i \} \). Furthermore, let \( \alpha_{kbi} \geq 0 \) denote the fraction of resources that BS \( b \) allocates to user \( k \) under pattern \( i \). Naturally we have \( \sum_{k \in K} \alpha_{kbi} \leq \pi_i, \forall b, \forall i \). Note that the user association is implicitly indicated by \( \alpha_{kbi} \), i.e., \( \alpha_{kbi} > 0 \) means user \( k \) is associated with BS \( b \) under pattern \( i \), while zero value of \( \alpha_{kbi} \) means that they are not connected. In this formulation, user \( k \) is allowed to be connected to multiple BSs.

Then we can express the average user rate after resource allocation as

\[
R_{\text{unrst}}^k = \sum_{i \in I} \sum_{b \in B} \alpha_{kbi}r_{kbi}
\]

where the superscript “unrst” is used to indicate that we refer to (2) as unrestricted user rate, in the sense that the user can have different multiple associations under different patterns.

To achieve the user rate promised in (2), user data should be available at multiple BSs and user equipment should have the ability to receive multiple data streams from different BSs simultaneously. In 3GPP LTE-Advanced standards, mobiles with inter-site carrier aggregation capability can be scheduled by multiple BSs over multiple frequencies \([27]\). Nevertheless, the unrestricted multiple associations introduce significant overheads and increase the system complexity. It is of both practical and theoretical interest to investigate the cardinality-constrained user association, where the maximum number of serving BSs for a given user is constrained. This restricted user rate is represented as follows:

\[
R_{\text{rst}}^k = \sum_{i \in I} \sum_{b \in B} a_{kb}\alpha_{kbi}r_{kbi}
\]

where \( a_{kb} \in \{0, 1\} \) is introduced to express the association explicitly, i.e., \( a_{kb} = 1 \) if user \( k \) is associated with BS \( b \); 0 otherwise. Thus, the number of user \( k \)'s serving BSs can be controlled by imposing the constraint \( \sum_{b \in B} a_{kb} = C_k \). If \( C_k = 1 \), we enforce unique association for user \( k \).

Remark 1: The ergodic rate is used in our system model because of two reasons. First, the adaptation of user association and multi-cell resource assignment is expected to perform over a large time scale, e.g., in minutes or hours. Too frequent changes of user association and multi-cell resource allocation will introduce severe overhead and delay issues and hence deteriorate the user experience. So it is impractical to adapt according to the fast fading channel state information (in milliseconds). Second, the ergodic rate formulation results in equal channel conditions among all frequency resources, which enables a unified treatment for either time or frequency allocation such that the user rate scales linearly with the assigned resources as shown in (2) and (3).
Remark 2: In this paper, we aim at developing strategies for long-term user association and multi-cell resource allocation based on the average channel (or more accurately, user rate) information. On top of this adaptation, each cell can perform individual channel-aware scheduling for all its associated users among the agreed spectrum in a more frequent manner to respond to fast fading channel fluctuations. Our design serves the first layer in this two-level adaptation.

III. UNRESTRICTED USER ASSOCIATION AND PATTERN SELECTION

In this section, we study the problem of joint user association and pattern selection without considering cardinality constraints on user association or number of patterns we can use. It will provide an ultimate limit on network performance and an optimal benchmark for comparison to other approaches where user association and pattern selection are restricted.

A. Problem formulation

Our objective is to maximize the long-term network utility, which is formulated as

$$\text{maximize } U = \sum_{k \in K} \omega_k \log(R_k)$$  \hspace{1cm} (4a)

subject to

$$R_k = \sum_{i \in I} \sum_{b \in B} \alpha_{kbi} r_{kbi}$$  \hspace{1cm} (4b)

$$\sum_{k \in K} \alpha_{kbi} \leq \pi_i, \forall b, \forall i$$  \hspace{1cm} (4c)

$$\sum_{i \in I} \pi_i = 1$$  \hspace{1cm} (4d)

$$\pi_i \geq 0, \alpha_{kbi} \geq 0$$  \hspace{1cm} (4e)

where a logarithmic utility function is used because it strikes a very good balance between network throughput and fairness among the users, and is a very common choice in existing studies (e.g., [1], [16], [17]), the weights $\omega_k$ provide a means for service differentiation. Note that our framework also applies to any other concave differentiable utility functions.

Remark 3: The formulation of (4) in [16] can be regarded as a special case of ours by letting reuse-1 be the only allowable pattern in our formulation. Our formulation introduces multi-pattern resource allocation as another degree of freedom for system design.

Remark 4: The investigation in [23] can also be casted into our framework. Specifically, by restricting the candidate patterns to two with partially shared resources, our formulation boils down to the optimization problem formulated in (5) in [23]. Being more general, our formulation can be used to investigate the case where each macro BS has different blank subframes.
Our formulation of (4) is a convex optimization problem since we maximize a concave function over a convex set. However, solving (4) is nontrivial because of the extremely high problem dimension in a reasonably-sized network. For example, consider a network with 15 cells, 90 users. The number of possible patterns (dimension of $\pi$) becomes $2^{15} - 1 = 32767$, and the dimension of variable $\alpha$ is then $90 \times 15 \times 32767 = 44,235,450$. Moreover, the number of constraints in (4c) is $15 \times 32767 = 491,505$. In this case, the general-purpose solver will fail due to the memory issue or unacceptable running time. In the following, we will present an efficient algorithm tailored to the problem at hand.

B. Some properties

For the ease of presentation, let us first define the set of serving BSs, the multi-associated user and the set of active patterns formally.

**Definition 1 (Set of serving BSs):** The serving BS for user $\bar{k}$ is the BS where user $\bar{k}$ is allocated nonzero fraction of resources over at least one pattern. The set of them is defined as $B_{\bar{k}} = \{b \in B : \alpha_{\bar{k}b} > 0\}$, where $\alpha_{\bar{k}b} = \sum_{i \in \mathcal{I}} \alpha_{\bar{k}bi}$.

**Definition 2 (Multi-associated user):** User $\bar{k}$ is referred to as a multi-associated user if it has more than one serving BS, i.e., $|B_{\bar{k}}| > 1$.

**Definition 3 (Set of active patterns):** The set of active patterns that user $\bar{k}$ gets from BS $\bar{b}$ is defined as $I_{\bar{k}\bar{b}}^\text{on} = \{i \in \mathcal{I} : \alpha_{\bar{k}bi} > 0\}$; The set of active patterns in the network is defined as $I_{\text{on}} = \{i \in \mathcal{I} : \pi_i > 0\}$.

The following Propositions 1 and 2 reveal properties of the optimal solution to problem (4), motivating the development of our algorithm.

**Proposition 1:** Among all optimal solutions to convex problem (4), we can always find one solution by activating at most $K + 1$ patterns in the network, i.e., $|I_{\text{on}}| \leq K + 1$, where $K$ is the number of users in the network.

**Proof:** By letting $\alpha_{kbi} = \pi_i \rho_{kbi}$, the problem of (4) can be equivalently rewritten as

$$\begin{align*}
\text{maximize } \quad & U = \sum_{k \in \mathcal{K}} \omega_k \log(R_k) \\
\text{subject to } \quad & R_k = \sum_{i \in \mathcal{I}} \sum_{b \in \mathcal{B}} \pi_i \rho_{kbi} r_{kbi} \\
& \sum_{k \in \mathcal{K}} \rho_{kbi} \leq 1, \forall b, \forall i \in I_{\text{on}} \\
& \sum_{i \in \mathcal{I}} \pi_i = 1 \\
& \pi_i \geq 0, \quad \rho_{kbi} \geq 0.
\end{align*}$$
Let us denote the user rate vector as \( \mathbf{R} = [R_1, \ldots, R_K]^T \). The achievable rate region defined by the constraints from (5b) to (5e) can be expressed as

\[
\mathcal{R} = \{ \mathbf{R} : \exists \mathbf{R}^i \in \mathcal{R}^i, \ \exists \pi \in \Pi, \ \mathbf{R} = \sum_i \pi_i \mathbf{R}^i \} = \text{conv}(\mathcal{R}^1, \ldots, \mathcal{R}^I) \tag{6}
\]

where \( \Pi = \{ \pi : \sum_i \pi_i = 1, \pi_i \geq 0 \} \), and

\[
\mathcal{R}^i = \{ \mathbf{R}^i = [R^i_1, \ldots, R^i_k, \ldots]^T : R^i_k = \sum_{b \in B} \rho_{kba} r_{kbi}, \ \sum_{k \in K} \rho_{kbi} \leq 1, \ \rho_{kbi} \geq 0 \} \tag{7}
\]

According to the Caratheodory’s theorem [28, Theorem 2.1.6], any point \( \mathbf{R} \in \mathcal{R} \subset \mathbb{R}_+^K \) can be expressed as the convex combination of at most \( K + 1 \) points in \( \bigcup_{i=1}^I \mathcal{R}^i \). Hence no more than \( K + 1 \) patterns are needed to achieve the same optimum as with any higher number of active patterns.

An immediate consequence of Proposition 1 is that although the number of all possible patterns grows exponentially with the number of cells in the network, we can allocate nonzero fraction of resources to only a small number of patterns to achieve the optimality. Note that resource allocation among the set of transmission modes has been studied in a TDMA context in [25], [29]. They conjectured that almost always only very few active transmission modes are needed as corroborated by their simulation results. Relying on the tool of Caratheodory’s theorem, we are able to bound the cardinality of the set of active patterns theoretically, which as a byproduct validates their conjecture.

Note that since the problem is not strictly convex, the optimal solution is not unique. How to pursuit the sparse solution in the sense of activating less patterns is not answered by the Proposition 1, which is the topic in the next subsection.

**Proposition 2:** In all optimal solutions to convex problem (4), the number of multi-associated users is at most \( B - 1 \), where \( B \) is the number of cells in the network.

**Proof:** The proof is based on the KKT conditions and provided in Appendix II.

The implication of Proposition 2 is remarkable, since it indicates as a result of optimal resource allocation and user association, most of users in the network will be associated with only one BS although we allow multi-association in our problem formulation. The number of multi-associated user is independent of the total number of users in the network, and it is upper bounded by the number of BSs.

**C. Proposed approach**

Our approach stems from Frank-Wolfe method, also known as the conditional gradient method [30]. As one of the simplest first-order methods that has been known since 1950s, Frank-Wolfe-type methods have recently re-gained interest in several areas, including machine learning particularly, mainly due to
its good scalability, and the crucial property of enabling sparse solutions [31]. We present our approach in the following Algorithm 1.

**Algorithm 1** Sparsity Pursuit for Joint User Association and Pattern Selection

1: Initialize iteration counter $t = 1$, tolerance $\epsilon > 0$; Choose $(\alpha^1, \pi^1) \in \mathcal{X}$;

2: repeat

3: Compute $(\bar{\alpha}, \bar{\pi}) = \arg \max_{(\alpha, \pi) \in \mathcal{X}} \langle \alpha, \nabla_\alpha U(\alpha^t) \rangle$;

4: Update $\alpha^{t+1} = \alpha^t + \gamma^t (\bar{\alpha} - \alpha^t)$, $\pi^{t+1} = \pi^t + \gamma^t (\bar{\pi} - \pi^t)$, where step size $\gamma^t \in [0, 1]$ is chosen by a warm-start Armijo rule;

5: $t = t + 1$;

6: until optimality certificate is less than $\epsilon$.

In Algorithm 1, we denote the feasible set defined by (4c), (4d) and (4e) as $\mathcal{X}$. The gradient matrix $\nabla_\alpha U(\alpha^t)$ is defined with the entries $[\nabla_\alpha U(\alpha^t)]_{kbi} = \frac{\partial U(\alpha)}{\partial \alpha_{kbi}}|_{\alpha = \alpha^t}$. The inner product $\langle \alpha, \nabla_\alpha U(\alpha^t) \rangle = \sum_{k=1}^{K} \sum_{b=1}^{B} \sum_{i=1}^{I} \alpha_{kbi} [\nabla_\alpha U(\alpha^t)]_{kbi}$. We explain the key elements of Algorithm 1 as follows.

1) **Solving the linear subproblem**: Solving the linear subproblem in line 3 is the most critical step for Algorithm 1. In particular, state-of-the-art linear programming solvers could fail due to the extremely high problem dimension that we are interested in. Fortunately, a closer look reveals that it has the following explicit analytic solution:

**Proposition 3**: The solution to the linear subproblem in Algorithm 1 is

$$\bar{\alpha}_{kbi} = \begin{cases} 1 & \text{if } i = \bar{i}, k = \bar{k}(b, \bar{i}), \forall b \\ 0 & \text{otherwise} \end{cases}$$

and

$$\bar{\pi}_i = \begin{cases} 1 & \text{if } i = \bar{i} \\ 0 & \text{otherwise} \end{cases}$$

where

$$\bar{k}(b, \bar{i}) = \arg \max_k [\nabla_\alpha U(\alpha^t)]_{kbi}$$

$$\bar{i} = \arg \max_i \sum_b [\nabla_\alpha U(\alpha^t)]_{k(b,i)bi}$$
Proof: The linear subproblem can be rewritten as the following inner-outer formulation:

\[
\begin{align*}
\max_{\pi_i \geq 0, \sum_i \pi_i = 1} & \quad \sum_{i=1}^{K} \sum_{b=1}^{B} \sum_{l=1}^{I} \alpha_{kbi} [\nabla_\alpha U(\alpha^l)]_{kbi} \\
\end{align*}
\]

(12)

It is clear that the inner problem is solved by each BS exclusively allocating maximum allowable resources to one user who benefits the most for each pattern, i.e.,

\[
\alpha_{kbi} = \begin{cases} 
\pi_i & \text{if } k = \bar{k}(b,i), \forall b, \forall i \\
0 & \text{otherwise}
\end{cases}
\]

(13)

where \(\bar{k}\) is expressed as (10). Substituting the solution of (13) back to (12), we arrive at the following problem:

\[
\max_{\pi_i \geq 0, \sum_i \pi_i = 1} \sum_{i=1}^{I} \sum_{b=1}^{B} [\nabla_\alpha U(\alpha^l)]_{\bar{k}(b,i)bi}
\]

(14)

which is solved by pooling all resources to one pattern. So we obtain the solutions of (11) and (9), hence (8).

Remark 5: Proposition 3 reveals that at most one new pattern is activated in each iteration. Hence, by initializing it with single pattern, Algorithm 1 has the potential to identify sparse solutions of using at most \(t\) patterns, where \(t\) is the iteration counter. Actually, as will been shown in the numerical results, the solution found by the Algorithm 1 has the number of active patterns close to \(K\), much less than \(t\) (meaning that no new patterns are activated for certain number of iterations).

2) Choosing the step size: The basic Armijio rule [30] is to set \(\gamma^t = \beta^m_t\), where \(\beta \in (0, 1)\) and \(m_t\) is the first nonnegative integer for which

\[
U(\alpha^{t+1}) \geq U(\alpha^t) + \kappa (\alpha^{t+1} - \alpha^t, \nabla_\alpha U(\alpha^t))
\]

(15)

with \(\kappa \in (0, 1)\) fixed. The idea is that we start with the initial step size 1 and continue to reduce to \(\beta, \beta^2, \ldots\), until the we find the largest \(\beta^{m_t}\) such that the \(\kappa\) improvement of utility function by its linear estimation is achieved.

We adopt a warm-start variant of the basic Armijio rule. In view of the fact that \(\gamma^{t-1}\) and \(\gamma^t\) may be similar, instead of starting from 1 every time we use \(\gamma^{t-1}\) as the initial guess and then either increase or decrease it in order to find the largest \(\beta^{m_t}\) satisfying (15). Specifically, we initially set \(\gamma^t = \gamma^{t-1}\); If the condition (15) is satisfied, we set \(\gamma^t = \min(\gamma^t/\beta, 1)\) until (15) does not hold or \(\gamma^t = 1\); Else repeatedly decrease \(\gamma^t\) by setting it to \(\beta \gamma^t\).
3) Initialization: In order to obtain sparse solution, we start the algorithm with single-BS association for each user and a single active pattern that gives the largest objective. Specifically, each user chooses the BS with largest $r_{kbi}$ for each pattern, resulting in the number of users at each BS for each pattern as $K_{bi}$. We allocate all resources to one pattern and distribute inside each BS uniformly, as

$$
\pi_i^1 = \begin{cases} 
1 & \text{if } i = \hat{i} \\
0 & \text{otherwise}
\end{cases}
$$

and

$$
\alpha_{kbi}^1 = \begin{cases} 
\frac{1}{K_{bi}} & \text{if } i = \hat{i} \text{ and } K_{k\hat{i}} \neq 0 \\
0 & \text{otherwise}
\end{cases}
$$

where

$$
\hat{i} = \arg\max_i \sum_k \omega_k \log \left( \sum_{b \in \{b: K_{bi} \neq 0\}} \frac{r_{kbi}}{K_{bi}} \right).
$$

4) Optimality certificate and convergence: We denote the $\alpha^*, \pi^*$ as the solution to problem (4), i.e., $U(\alpha^*) \geq U(\alpha), \forall (\alpha, \pi) \in \mathcal{X}$. The $\epsilon$-optimal solution set is defined as

$$
S^*_\epsilon = \{ (\alpha, \pi) \in \mathcal{X} : U(\alpha^*) - U(\alpha) \leq \epsilon \}.
$$

Due to the concavity of the objective function, we have

$$
U(\alpha^*) - U(\alpha^t) \leq \langle \nabla_\alpha U(\alpha^t), \alpha^* - \alpha^t \rangle \leq \maximize_{(\alpha, \pi) \in \mathcal{X}} \langle \nabla_\alpha U(\alpha^t), \alpha - \alpha^t \rangle
$$

where the second inequality follows because $\alpha^*$ is a feasible point.

Let us define the optimality gap function as

$$
g(\alpha^t) = \maximize_{(\alpha, \pi) \in \mathcal{X}} \langle \nabla_\alpha U(\alpha^t), \alpha - \alpha^t \rangle = \langle \nabla_\alpha U(\alpha^t), \alpha^* - \alpha^t \rangle.
$$

Hence, the value of $g(\alpha^t)$ can be easily obtained as a by-product of every iteration of Algorithm 1. If the current iteration satisfies $g(\alpha^t) \leq \epsilon$, it is guaranteed for $\alpha^t, \pi^t$ being $\epsilon$-optimal (i.e., $(\alpha^t, \pi^t) \in S^*_\epsilon$).

In order to use $g(\alpha^t) \leq \epsilon$ as the stopping criterion, we need to make sure that $g(\alpha) \to 0$ if $(\alpha, \pi) \to (\alpha^*, \pi^*)$. According to optimality condition of a differentiable concave function, the optimal solution should satisfy [32]

$$
\langle \nabla_\alpha U(\alpha^*), \alpha - \alpha^* \rangle \leq 0, \forall (\alpha, \pi) \in \mathcal{X}
$$

which means

$$
\maximize_{(\alpha, \pi) \in \mathcal{X}} \langle \nabla_\alpha U(\alpha^*), \alpha - \alpha^* \rangle = g(\alpha^*) = 0.
$$

As a result, we have the following convergence result:
Proposition 4: \((\alpha^t, \pi^t)\) generated by Algorithm 1 converges to \(\epsilon\)-optimal solution.

Proof: According to (21), (22) and Proposition 2.2.1 in [30], \(g(\alpha^t) \to 0\) as \(t \to \infty\). If we stop the algorithm after a finite number of steps and condition \(g(\alpha^t) \leq \epsilon\) is met, the solution is in the \(\epsilon\)-optimal solution set according to the definition (18).

Remark 6: Algorithm 1 provides a trade-off between the optimality quality and the number of active patterns in the final solution. Specifically, larger number of iterations is needed if we require a more accurate solution with a smaller \(\epsilon\). Consequently, the number of active patterns in the solution may increase because a new pattern could be added at every iteration. A notable feature of the algorithm is that whenever we stop the algorithm, we know how far the current solution is from the optimum.

IV. CARDINALITY-CONSTRAINED USER ASSOCIATION AND PATTERN SELECTION

In the previous section, user can have multiple associations and the associations can be different under different patterns. We now focus on more practical consideration that the number of serving BSs for a given user is restricted and the set of serving BSs is the same under all patterns for any given user.

A. Problem formulation and approach

In order to constrain the number of serving BSs for each user, we introduce a new association matrix variable \(a\) and formulate the utility maximization problem as follows:

\[
\begin{align*}
\text{maximize}_{\alpha, \pi, a} \quad & U = \sum_{k \in K} \omega_k \log(R_k) \\
\text{subject to} \quad & R_k = \sum_{i \in I} \sum_{b \in B} a_{kb} \alpha_{kbi} r_{kbi} \\
& \sum_{k \in K} a_{kb} \alpha_{kbi} \leq \pi_i, \forall b, \forall i \\
& \sum_{b \in B} a_{kb} = C_k, \forall k \\
& a_{kb} \in \{0, 1\} \\
& \sum_{i \in I} \pi_i = 1 \\
& \pi_i \geq 0, \quad \alpha_{kbi} \geq 0
\end{align*}
\]

where \(C_k\) is the number of serving BSs that user \(k\) is allowed to associate with.

The problem of (23) is a mixed-integer nonconvex problem, hence difficult to solve. Our approach is to leverage Algorithm 1 proposed in the previous section. Specifically, we group the variables into two
blocks, \((\alpha, \pi)\) and \(\alpha\), and optimize them alternatively with the other block fixed as shown in Algorithm 2. The details and interesting characteristics of Algorithm 2 are provided in the following Section IV-B.

**Algorithm 2** Cardinality-Constrained User Association and Pattern Selection

1: **Initialization**: \(a_{kb} = 1, \forall k, \forall b\);
2: **repeat**
3: Solve (23) for fixed \(\alpha\) by Algorithm 1;
4: Solve (23) for fixed \((\alpha, \pi)\);
5: **until** objective function cannot be increased.

---

**B. Algorithm details and characteristics**

1) **Pattern allocation for fixed user association**: We first show that the problem (23) with fixed association \(\alpha\) can be solved exactly by Algorithm 1. By defining \(\hat{r}_{kbi} = a_{kb}r_{kbi}\) and \(\mathcal{K}_b = \{k \in \mathcal{K} : a_{kb} = 1\}\), we rewrite (23) to the following problem

\[
\begin{align*}
\text{maximize} & \quad U = \sum_{k \in \mathcal{K}} \omega_k \log(R_k) \\
\text{subject to} & \quad R_k = \sum_{i \in I} \sum_{b \in \mathcal{B}} \alpha_{kbi} \hat{r}_{kbi} \\
& \quad \sum_{k \in \mathcal{K}_b} \alpha_{kbi} \leq \pi_i, \forall b, \forall i \\
& \quad \sum_{i \in I} \pi_i = 1 \\
& \quad \pi_i \geq 0, \quad \alpha_{kbi} \geq 0
\end{align*}
\]

which is almost the same as problem (4) with the only two differences. First, the summation in the constraint (24c) is over \(\mathcal{K}_b\). Second, the problem (24) is solved when the effective rate \(\hat{r}_{kbi}\), instead of \(r_{kbi}\) are given. As (4), problem (24) is also convex and has the same structure. Hence Algorithm 1 can be directly applied to solve the problem to the desired accuracy. Interestingly, the maximization operation in (10) can be taken over \(\mathcal{K}\) without damaging the optimality (see following Proposition 5).

The following proposition states an interesting property of the solution obtained by applying Algorithm 1 to problem (24), which we refer to as universal feasibility.

**Proposition 5**: Let \((\alpha^\circ, \pi^\circ)\) denotes the solution to problem (24) by using Algorithm 1. Then we have

\[\sum_{k \in \mathcal{K}} \alpha_{kbi}^\circ \leq \pi_i^\circ, \forall b, \forall i.\]
Proof: According to Algorithm 1, the final solution is the convex combination of extreme points identified by solving the linear subproblem. Strictly speaking, when applying Algorithm 1 to problem (24) the maximization operation in (10) is over $K_b$ due to the constraints in (24c). However, since
\[
\nabla \alpha U(\alpha)_{kbi} = \frac{\omega_k \hat{r}_{kbi}}{R_k} = \frac{\omega_k a_{kbi} r_{kbi}}{R_k} = \omega_k r_{kbi} R_k = \omega_k a_{kbi} r_{kbi}
\]

obtain minimum values (which are zeros) at $k \notin K_b$, the result of (10) will have $k(b, i) \in K_b$ even the maximization is over $K$, provided $K_b$ is not empty and BS $b$ is not muted under pattern $i$ (i.e., $\exists k, r_{kbi} \neq 0$). Hence $\{\alpha_{kbi}, k \notin K_b, \forall b, \forall i\}$ are zeros in the final solution. This proves the proposition. Note that zero values of $\{\alpha_{kbi}, k \notin K_b, \forall b, \forall i\}$ will not degrade the optimality of the solution, because $\hat{r}_{kbi}$ are also zeros at those positions. In the case of empty $K_b$ or BS $b$ is muted under pattern $i$, we have $\nabla \alpha U(\alpha)_{kbi} = 0, \forall k \in K$. Then the maximization in (10) will give an arbitrary $\bar{k}$. The above proposition still holds.

Remark 7: The universal feasibility identified by Proposition 5 will significantly simplify the calculation of user association update as given in the following subsection, by removing the coupling constraints.

2) User association update: After we define $R_{kb} = \sum_i \alpha_{kbi} r_{kbi}$, the problem (23) with fixed $(\alpha, \pi)$ can be rewritten as

\[
\text{maximize} \quad U = \sum_{k \in K} \omega_k \log(R_k) \quad \text{(25a)}
\]

\[
\text{subject to} \quad R_k = \sum_{b \in B} a_{kb} R_{kb} \quad \text{(25b)}
\]

\[
\sum_{b \in B} a_{kb} = C_k, \forall k \quad \text{(25c)}
\]

\[
a_{kb} \in \{0, 1\}. \quad \text{(25d)}
\]

Note that constraints in (24c) have been dropped because they are automatically satisfied due to the universal feasibility property proved by Proposition 5.

It is easily seen that problem (25) can be decoupled into $K$ problems, which can be solved exactly by each user choosing the dominant $C_k$ BSs that give the largest $R_{kb}$, i.e.,

\[
a_{\hat{c}k} = \begin{cases} 1 & \text{if } b \in b^\circ(k) \\ 0 & \text{otherwise} \end{cases} \quad \text{where } b^\circ(k) = \{b \in B : R_{kb} \geq R_{k(C_k)}\}
\]

with $R_{k(C_k)}$ denoting $C_k$-th largest element in $(R_{k1}, \cdots, R_{kB})$.

3) Initialization and convergence: A natural choice of initialization is the unrestricted user association, i.e., $a_{kb} = 1, \forall k, \forall b$. Although it is not a feasible $\alpha$ in terms of the cardinality constraints, it produces the upper bound after applying Algorithm 1, and then becomes feasible after one iteration of Algorithm 2. It beats random initialization significantly in our experiments.

August 22, 2014 DRAFT
Since both the pattern allocation update and user association update in Algorithm 2 maximize the same utility function, the overall algorithm produces nondecreasing objective function values. The whole algorithm is guaranteed to converge because the utility function is finite. Although the converged solution is not necessarily global optimal, we bound the performance loss by comparing it to the upper bound obtained by unrestricted user association.

V. PERFORMANCE EVALUATION AND DISCUSSIONS

A. Simulation scenarios and parameter setting

We consider a network consisting of 3 macro cells, each of which contains 4 randomly dropped pico cells as shown in Fig. 2. The cells are labelled as

![Cells Labelling](image)

The parameters for propagation modelling and simulations follow the suggestions in 3GPP evaluation methodology [33], and summarized in Table I. In the area under consideration, we choose user equipment (UE) density of around 420 and 225 active UEs/sq-km (corresponding to dense urban and urban environment [34]), resulting in $K = 90$ and 50 respectively. All UEs have unit weights ($\omega_k = 1$). In
TABLE I
NETWORK PARAMETERS.

| Parameter                  | Description                                      |
|----------------------------|--------------------------------------------------|
| bandwidth                  | 10 MHz                                           |
| Macro total Tx power       | 46 dBm                                           |
| Pico total Tx power        | 30 dBm                                           |
| Macro antenna gain         | 15 dB                                            |
| Pico antenna gain          | 5 dB                                             |
| Macro path loss            | $128.1 + 37.6 \log_{10}(R)$                      |
| Pico path loss             | $140.7 + 36.7 \log_{10}(R)$                      |
| Penetration loss           | 20 dB                                            |
| Shadowing std. dev.        | 8 dB(macro), 10 dB(pico)                         |
| Shadowing corr. distance   | 25 m                                             |
| Macrocell shadowing corr.  | 1 between cells                                  |
| Picocell shadowing corr.   | 0.5 between cells                                |
| Fading model               | No fast fading                                   |
| Min. macro(pico)-UE dist. | 35 m (10 m)                                      |
| Min. macro(pico)-pico dist.| 75 m (40 m)                                      |
| Noise density and noise figure | -174 dBm/Hz, 9dB                               |

simulation, the total number of users are randomly dropped in the network, and we average over 5 drops of users and pico locations for each UE density.

There are only two parameters regarding the step size rule, $\gamma^0$ and $\beta$, which need to be set before our algorithms are run, and the performance of our algorithms is insensitive to their values. In the simulation, we set initial step size $\gamma^0 = 10^{-4}$ and $\beta = 0.8$. For the optimality tolerance, we set $\epsilon = 1$ for low-UE-density case and $\epsilon = 2$ for the high-UE-density case.

B. Feature pattern identification by Algorithm 1

As the number of all possible patterns in the network grows exponentially with the number of cells, it is necessary to identify the most important patterns and allocate resources only to these feature patterns in order to reduce the complexity of the algorithm. As described in Section III-C, the proposed Algorithm 1 has sparsity-pursuit capability to activate only a small number of patterns in the final solution. In this subsection, we focus on the feature pattern identification with the aid of Algorithm 1.

For a macro-only homogeneous network, [25] suggested a guideline for candidate pattern selection in
the TDMA macro networks by choosing only two kinds of patterns: 1) reuse-1: all BSs are active and 2) all neighboring BSs except one are active. However, this guideline cannot be applied to the HetNet that we are investigating, because the neighboring cells are not well-defined due to the overlaid deployment. Moreover, the significantly larger number of small cells also makes this approach produce too many candidate patterns.

In another study [1], several strategies of selecting reuse patterns in HetNet have been proposed by intuition and their performance have been compared. Among them, orthogonal and partially shared deployments with reuse-1 among pico BSs perform significantly better than co-channel deployment. However, the question whether there exists better strategies remains open.

We answer this question by considering all possible patterns in the tested network, and rely on the sparsity-pursuit capability of Algorithm 1 to identify the feature patterns. Our approach provides a systematic way to find the best reuse strategies. Table II lists the number of active patterns after Algorithm 1 converges to $\epsilon$-optimal solution. Recall that pattern $i$ is referred to as active if $\pi_i > 0$ in the final solution. As shown in Table II, the proposed Algorithm 1 is very effective in finding the sparse solution, in the sense that only around $K$ patterns are actually used out of $(2^{15} - 1)$ patterns. For a complete characterization of the behavior of Algorithm 1, the numbers of iterations executed for different random drops are also provided in Table III where we observe larger values than the numbers of active patterns given in Table II. This implies that some iterations are simply carried out on previously-activated patterns.

Obviously, the set of active patterns generated by Algorithm 1 are different in each drop, varying in
accordance to the user distribution, pico distribution, fading environment, etc. After examining all the results we tested, we propose the following general guideline for feature pattern selection in a HetNet without resorting to the Algorithm 1.

- All macros are OFF, and all picos are ON.
- One macro is ON among the adjacent three macros, and all picos are ON except those in the active macro cells.

The principle can be summarized as macro-OFF-pico-ON policy. One advantage of this policy is that the resulting set of candidate patterns is very small. For example, in our tested network, there are only 4 candidate patterns needed by applying this guideline. Note that in our preliminary work [26] a similar selection criterion was developed for a different framework via heuristic Tabu-search approach, where each user is allowed to associate with only one BS and each BS serves multiple users by round-robin scheduler, allocating all its sub-channels to a single user at a given time. In contrast, this paper derives the guideline by analyzing the optimal solution of a more general formulation. More importantly, the Algorithm 1 developed in this paper can be used to calculate the optimal solution by considering all patterns, benchmarking the sub-optimal/low-complexity strategies where we restrict the candidate patterns to the pre-selected ones. A detailed comparison is provided in the next subsection.

C. Comparing various strategies

In this subsection, we illustrate how to use our framework as a unified way to compare various existing user association and resource allocation schemes.

1) Comparative schemes: The six schemes that we compare are as follows.

- **Optimal Unrestricted Scheme (Opt)**: We consider all \((2^{15} - 1)\) patterns in the tested network as the candidate patterns, and no cardinality restriction is imposed on user association. This is the only scheme in comparison where multi-BS association is allowed. We optimize the user association and resource allocation by Algorithm 1 proposed in this paper.

- **Preselected Feature Patterns (Fea-Pattern)**: The candidate patterns are restricted to four preselected ones according to our guideline proposed in Section V-B. Specifically, if we denote a pattern by the set of muted BSs under this pattern, these four patterns are \(\{1, 2, 3\}\), \(\{2, 3, 4, 5, 6, 7\}\), \(\{1, 3, 8, 9, 10, 11\}\), and \(\{1, 2, 12, 13, 14, 15\}\). The joint optimization of user association and resource allocation is solved by Algorithm 2 proposed in this paper.

- **Orthogonal Deployment with Reuse-1 (OD-1)**: This is one of the strategies studied in [1], where the macro layer and pico layer are deployed on the different frequency channels, so there is no inter-
layer interference. The intra-layer interference is handled by a reuse factor. In this OD-1, reuse-1 is simply used to allow maximum bandwidth at each BS. In [1], the channel allocation between macro and pico layers was solved, together with user association, by exhaustive search, in view of the fact that the set of channels in the system is discrete and finite. Here we study this strategy by using our framework. In particular, this OD-1 deployment can be regarded as restricting the candidate patterns in our framework to two, which mute all macro and all pico cells respectively. Namely, \(\{1, 2, 3\}\) and \(\{4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15\}\) are the only two allowed patterns. The joint user association and frequency allocation over these two patterns can be solved by our Algorithm 2.

- **Orthogonal Deployment with Pico Reuse-3 (OD-3)**: This is another strategy investigated in [1], which is similar to OD-1. The only difference is that the pico cells share the frequency channels allocated to the pico layer by a reuse factor of three. To study OD-3 using our framework, we simply restrict the candidate patterns to the following four by muting: \(\{4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15\}\), \(\{1, 2, 3, 4, 6, 7, 9, 10, 12, 13, 15\}\), and \(\{1, 2, 3, 4, 5, 7, 8, 10, 11, 13, 14\}\), respectively. Algorithm 2 is also used to optimize the user association and frequency allocation.

- **Synchronous Blank Subframes in Macro Tier (Macro ABS)**: This is the time domain enhanced inter-cell interference coordination (eICIC) scheme studied in [23]. We can easily investigate this strategy using our Algorithm 2, by defining two candidate patterns, muting \(\{\emptyset\}\) and \(\{1,2,3\}\), respectively.

- **Reuse-1**: By assuming reuse-1, user association problem has been studied in [16], which is a special case of our formulation by restricting candidate pattern to one single pattern that activating all cells (i.e., muting \(\{\emptyset\}\)). The Algorithm 2 proposed in this paper can be used to solved the resulting optimization problem.

We summarize how we study various schemes using our framework in Table IV.

2) **Performance metrics**: The following performance metrics are considered in our comparison.

- Geometric mean of user throughput, defined as \(\sqrt[k]{\prod_{k=1}^{K} R_k}\). Note that maximizing the geometric
mean throughput is equivalent to maximizing our log utility with $\omega_k = 1$. 

- Total sum rate of the system, defined as $\sum_{k=1}^{K} R_k$.
- Cumulative Distribution Function (CDF) of the user throughput.

The reason why we are interested in more metrics than objective function is that both system sum rate and CDF of user throughput are key performance indicators (KPIs) for network operators. We want to show how different user association and resource allocation schemes impact on these KPIs.

3) Results: Fig. 3 reports the geometric mean of user rate and total system sum rate for all schemes in comparison. As shown, our proposed Opt indeed provides an optimal benchmark, since it optimizes the utility function over all possible patterns. Fea-Pattern scheme that relies on the proposed practical guideline to identify the feature patterns is very effective, achieving 89% and 91% of the benchmark in terms of the geometric mean rate in 50-UE and 90-UE cases respectively, and 92% and 94% in terms of the sum rate. This is remarkable because the optimization is only performed over the four selected candidate patterns, requiring much less computational complexity in comparison to Opt. It performs significantly better than the existing strategies, namely, OD-1, OD-3, Macro ABS, and Reuse-1.

Among those existing strategies, Macro ABS performs the closest to our Fea-Pattern, immediately followed by OD-1. This is because these two existing methods happen to share the similar spirit of our pattern selection guideline by treating the umbrella macro as the dominant interferer to all the pico cells within its coverage and defining one pattern to mute this strongest interferer. The advantage of Macro ABS over OD-1 indicates that partially sharing resources between macro and pico tiers is more efficient than orthogonal deployment.
OD-3 suggests to deploy reuse-3 among all picos, which turns out to be inefficient as evident in Fig. 3. The reason is that only limited interference among pico BSs due to the low transmit power. Including pico reuse-3 patterns reduces the available resources at each pico BS and hence damages the system sum rate as shown in the right panel of Fig. 3.

As expected, Reuse-1 obtains the worst geometric mean rate since there is no mechanism for inter-cell interference management. In terms of the sum rate, however, it performs better than OD-3. This is because the cell-center users who are less affected by inter-cell interference can benefit from the access to full resources provided by Reuse-1. We make this point clear by providing the CDF of user throughputs in the next.

Figs. 4 and 5 show the empirical CDF of user throughputs for 50 UEs and 90 UEs in the network, respectively. Being consistent with results shown in Fig. 3, the proposed Fea-Pattern produces the closest curve to Opt. Both Macro ABS and OD-1 are less effective in reducing the interference compared to Fea-Pattern, resulting in performance loss for the cell-edge users (5th percentile throughput). Moreover, OD-1 is not as good as Macro ABS in utilizing the resources, causing further degradation for cell-center users (95th percentile throughput) compared to Macro ABS. As for OD-3, the large gap between OD-3 and all the other schemes with respect to 95th percentile throughput reveals that pico reuse-3 deployment in OD-3 is a waste of system resources, causing notable degradation in sum rate as shown in the right panel of Fig. 3. When comparing Fig. 4 and Fig. 5 we notice that the performance gaps among different schemes reduce when number of users in the network increases. This is because less resources are available for individual users as the number of users increases while the total resources are fixed in the network. So the gains achieved by optimizing the resource allocation decrease.

D. Comparison with RE association rule

Range expansion (RE) techniques have recently been discussed in 3GPP as a simple association rule to balance the load in HetNets. In this subsection, we study the impact of this simple association on the performance if resource allocation can be optimized, using our framework. Specifically, when the user association is fixed, the resource allocation can be optimized by solving problem (24) using Algorithm 1 as described in Section V-B1. In the evaluation, we set the macro bias to zero, and the same bias for all the pico BSs, choosing from 0, 5, 10, 15, 20 and 25 dB. The UE is associated with the cell with largest value of downlink received power plus bias.

Fig. 6 provides the comparison of Fea-Pattern scheme with other schemes of fixed RE association, where the resource allocation is performed over the same four feature patterns. As seen, the conventional
Fig. 4. CDF of user throughputs with different schemes for 50 UEs in the network

Fig. 5. CDF of user throughputs with different schemes for 90 UEs in the network

association scheme without bias can only obtain roughly 50% of the geometric mean rate that is achieved by the joint optimization of the Fea-Pattern scheme in both 50-UE and 90-UE cases. However, by tuning the bias value to optimum (20 dB in the tested case), the simple RE association method with optimized resource allocation achieves 90% of the geometric mean rate offered by Fea-Pattern. The small gap exists
because all pico BSs have to select the same bias value.

Fig. 7 shows the similar comparison, but the resource allocation is restricted to single reuse-1 pattern. In this case, the resource allocation boils down to intra-cell resource distribution. As shown, the performance
loss caused by the conventional association without bias is significantly less than that in Fig. 6. This is because reuse-1 pattern suffers from lack of inter-cell interference management, limiting the potential of joint optimization. Like in Fig. 6, the RE association with optimal bias (5 dB in Fig. 7) can achieve 95% performance of the joint optimization. Another interesting observation is that the optimal bias value in Fig. 7 is smaller than that in Fig. 6, indicating that with mechanisms to combat the inter-cell interference pico cells can use more aggressive bias value to incorporate more users.

Although Figs. 6 and 7 demonstrate the effectiveness of the RE association rule, it is a nontrivial task to select the optimal bias value for this method to work. Unlike the proposed schemes where we have a systematic way to optimize the user association with resource allocation, the RE association rule often requires an exhaustive search (as we did in Figs. 6 and 7) or some heuristics. The fact that optimal bias values are different in Figs. 6 and 7 also reveals the coupling effect of the resource allocation and user association.

VI. CONCLUSION

In this paper, we have studied the joint user association and interference management in heterogeneous networks. We treated the multi-cell resource allocation as resource partitioning among multiple reuse patterns, and developed efficient algorithms to optimize the multi-cell multi-user resource assignment together with user association. For the case of unrestricted user association, we formulated a convex problem and proposed a efficient sparsity-pursuit algorithm to find the optimal sparse solution where only a small number of patterns are activated. The result provides an optimal benchmark for comparison. On the other hand, the formulated problem becomes nonconvex and combinatorial for the case of restricted user association. By exploiting the problem structure, we were able to derive an efficient iterative algorithm to obtain solutions close to the optimal benchmark.

An important observation is that although the number of all possible patterns grows exponentially with the number of cells in the network, most of the patterns are not used thanks to the sparsity pursuit capability of the proposed algorithm. This motivated us to restrict the candidate patterns to a set of pre-defined feature patterns in order to reduce the computation efforts of the resource allocation algorithm. By analyzing the active patterns resulting from the sparsity-pursuit algorithm, we developed practical guideline to select the feature patterns in a HetNet.

Our framework enabled us to compare a wide range of user association and resource allocation strategies in a unified view. Our results indicate that existing criteria for feature pattern selection result in large performance loss in comparison to the optimal benchmark. However, the feature patterns identified by
our guideline significantly improves the existing strategies. We also compared our joint optimization to
the strategies where the user association is performed separately according to simple range expansion
rules. We observed that the RE rule achieves only minor performance loss, provided the bias value is
chosen optimally and the resource allocation is optimized. The fact that the optimal bias values should
be set differently in alignment with different resource allocation strategies reveals the coupling effect of
both elements.

APPENDIX

PROOF OF PROPOSITION 2

Without loss of generality, we set \( \omega_k = 1, \forall k \). We define the Lagrangian associated with problem \( (4) \) as

\[
L(\alpha, \pi, \mu, \sigma) = \sum_k \log(R_k) + \sum_i \sum_b \mu_b(\pi_i - \sum_k \alpha_{kbi}) + \sigma(1 - \sum_i \pi_i),
\]

where \( \mu \) and \( \sigma \) are Lagrange multipliers associated with \( (4c) \) and \( (4d) \), respectively, the nonnegative
constraints in \( (4e) \) are considered implicitly.

The optimal solution should satisfy the following KKT conditions:

\[
\frac{r_{kbi}}{R_k} - \mu_{bi} = 0, \quad \text{if } \alpha_{kbi} > 0 \quad (27a)
\]

\[
\sum_b \mu_{bi} = \sigma, \quad \text{if } \pi_i > 0 \quad (27b)
\]

\[
\sum_k \alpha_{kbi} \leq \pi_i, \quad \mu_{bi}(\sum_k \alpha_{kbi} - \pi_i) = 0 \quad (27c)
\]

\[
\mu_{bi} \geq 0, \quad \alpha_{kbi} \geq 0, \quad \pi_i \geq 0, \quad \sum_i \pi_i = 1 \quad (27d)
\]

Assume that user \( \bar{k} \) is associated with two BSs, \( b \) and \( b' \). According to \( (27a) \) we have

\[
\frac{r_{kbi}}{R_k} = \mu_{bi}, \quad \forall i \in T_{kb}^{un} \quad (28)
\]

\[
\frac{r_{kbi}}{R_k} = \mu_{bj}, \quad \forall j \in T_{kb'}^{on} \quad (29)
\]

For the ease of notation, we define \( r_{kb} = \sum_{i \in T_{kb}^{un}} r_{kbi} \) and \( \mu_b = \sum_{i \in T_{kb}^{un}} \mu_{bi} \). From \( (28) \) and \( (29) \), we
obtain

\[
\frac{r_{kb}}{r_{kb'}} = \frac{\mu_b}{\mu_{b'}}. \quad (30)
\]
To show the number of multiple-associated users is upper bounded by $B - 1$, we follow the same argument as in [35] using a bipartite graph representation (BGR), where the multiple-associated users and BSs are denoted by nodes, and a edge is placed between a user and a BS to represent the association. We next show that the BGR contains a loop with zero probability.

Suppose there exists a loop in BGR, as shown in Fig. 8. Then a sequence of nodes $k_1, b_1, k_2, b_2, \ldots, k_n, b_n, k_1$ exists, where nodes are different otherwise we can find a smaller loop inside. According to (30), the above loop implies

$$
\frac{r_{k_1b_n}}{r_{k_1b_1}}, \frac{r_{k_2b_1}}{r_{k_2b_2}}, \ldots, \frac{r_{k_nb_{n-1}}}{r_{k_nb_n}} = \frac{\mu_{b_n}}{\mu_{b_1}}, \frac{\mu_{b_1}}{\mu_{b_2}}, \ldots, \frac{\mu_{b_{n-1}}}{\mu_{b_n}} = 1,
$$

which is obtained with zero probability. This is because $\frac{r_{k_1b_n}}{r_{k_1b_1}}, \frac{r_{k_2b_1}}{r_{k_2b_2}}, \ldots, \frac{r_{k_nb_{n-1}}}{r_{k_nb_n}}$ is a function of independent continuous random variables, resulting in a continuous random variable itself. The probability of equaling a constant is zero.

With this in mind, we construct the BGR by adding one multi-associated user at one time. It is clear that the number of multi-associated users added can not be greater than the number of BSs, otherwise the bipartite graph can not be constructed without a loop, which proves the proposition.

![Fig. 8. A loop in bipartite graph representation.](image)

REFERENCES

[1] D. Fooladivanda and C. Rosenberg, “Joint resource allocation and user association for heterogeneous wireless cellular networks,” IEEE Trans. Wireless Commun., vol. 12, no. 1, pp. 248–257, 2013.

[2] G. Boudreau, J. Panicker, N. Guo, R. Chang, N. Wang, and S. Vrzic, “Interference coordination and cancellation for 4G networks,” IEEE Commun. Mag., vol. 47, no. 4, pp. 74–81, 2009.

[3] N. Saquib, E. Hossain, and D. I. Kim, “Fractional frequency reuse for interference management in LTE-advanced hetnets,” IEEE Wireless Commun. Mag., vol. 20, no. 2, pp. 113–122, 2013.

[4] S.-E. Elayoutobi and B. Fourestie, “On frequency allocation in 3G LTE systems,” in Personal, Indoor and Mobile Radio Communications, 2006 IEEE 17th International Symposium on, 2006, pp. 1–5.
[5] B. Holfeld, A. Relitz, T. Wirth, and E. Jorswieck, “Joint multicell subchannel assignment with interference control and resource fairness in multiband OFDMA cellular networks,” in Dynamic Spectrum Access Networks (DYSPAN), 2014 IEEE International Symposium on, 2014, pp. 467–476. [Online].

[6] D. Lopez-Perez, X. Chu, and J. Zhang, “Dynamic downlink frequency and power allocation in OFDMA cellular networks,” IEEE Trans. Commun., vol. 60, no. 10, pp. 2904–2914, 2012.

[7] R. Chang, Z. Tao, J. Zhang, and C.-C. Ku, “A graph approach to dynamic fractional frequency reuse (ffr) in multi-cell OFDMA networks,” in Communications, 2009. ICC ’09. IEEE International Conference on, 2009, pp. 1–6.

[8] T. Novlan, J. Andrews, I. Sohn, R. Ganti, and A. Ghosh, “Comparison of fractional frequency reuse approaches in the OFDMA cellular downlink,” in Global Telecommunications Conference (GLOBECOM 2010), 2010 IEEE, 2010, pp. 1–5.

[9] A. Dotzler, W. Utschick, and G. Dietl, “Fractional reuse partitioning for MIMO networks,” in Global Telecommunications Conference (GLOBECOM 2010), 2010 IEEE, 2010, pp. 1–5.

[10] S. Ali and V. Leung, “Dynamic frequency allocation in fractional frequency reused OFDMA networks,” IEEE Trans. Wireless Commun., vol. 8, no. 8, pp. 4286–4295, 2009.

[11] G. R1-051341, “Flexible fractional frequency reuse approach,” in 3GPP TSG RAN WG1 Meeting, Seoul, Korea, 2005.

[12] R. Yates and C.-Y. Huang, “Integrated power control and base station assignment,” IEEE Trans. Veh. Technol., vol. 44, no. 3, pp. 638–644, 1995.

[13] S. Hanly, “An algorithm for combined cell-site selection and power control to maximize cellular spread spectrum capacity,” IEEE J. Sel. Areas Commun., vol. 13, no. 7, pp. 1332–1340, 1995.

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

[15] 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, 2013.

[16] 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, 2013.

[17] K. Shen and W. Yu, “Downlink cell association optimization for heterogeneous networks via dual coordinate descent,” in Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, 2013, pp. 4779–4783.

[18] R. Madan, J. Borran, A. Sampath, N. Bhushan, A. Khandekar, and T. Ji, “Cell association and interference coordination in heterogeneous lte-a cellular networks,” IEEE Journal on Selected Areas in Communications, vol. 28, no. 9, pp. 1479–1489, 2010.

[19] I. Guvenc, “Capacity and fairness analysis of heterogeneous networks with range expansion and interference coordination,” IEEE Communications Letters, vol. 15, no. 10, pp. 1084–1087, 2011.

[20] I. Guvenc, M.-R. Jeong, I. Demirdogen, B. Kecicioglu, and F. Watanabe, “Range expansion and inter-cell interference coordination (icic) for picocell networks,” in Proc. IEEE Vehicular Technology Conf. (VTC Fall), 2011, pp. 1–6.

[21] K. Son, S. Chong, and G. Veciana, “Dynamic association for load balancing and interference avoidance in multi-cell networks,” IEEE Trans. Wireless Commun., vol. 8, no. 7, pp. 3566–3576, 2009.

[22] Q. Kuang, J. Speidel, and H. Droste, “Joint base-station association, channel assignment, beamforming and power control in heterogeneous networks,” in Vehicular Technology Conference (VTC Spring), 2012 IEEE 75th, 2012, pp. 1–5.

[23] Q. Ye, M. Al-Shalash, C. Caramanis, and J. G. Andrews, “On/off macrocells and load balancing in heterogeneous cellular networks,” in Global Communications Conference (GLOBECOM), 2013 IEEE, 2013, pp. 3814–3819.
[24] Y. Jin and L. Qiu, “Joint user association and interference coordination in heterogeneous cellular networks,” IEEE Commun. Lett., vol. 17, no. 12, pp. 2296–2299, 2013.
[25] K. Son, Y. Yi, and S. Chong, “Utility-optimal multi-pattern reuse in multi-cell networks,” IEEE Trans. Wireless Commun., vol. 10, no. 1, pp. 142–153, 2011.
[26] Q. Kuang, “Joint user association and reuse pattern selection in heterogeneous networks.” accepted in ISWCS 2014. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=5783991
[27] K. Pedersen, F. Frederiksen, C. Rosa, H. Nguyen, L. Garcia, and Y. Wang, “Carrier aggregation for LTE-advanced: functionality and performance aspects,” IEEE Commun. Mag., vol. 49, no. 6, pp. 89–95, 2011.
[28] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, Nonlinear programming: theory and algorithms, 3rd ed. New York: Wiley-Interscience, 2006.
[29] C. Raman, R. Yates, and N. B. Mandayam, “Scheduling variable rate links via a spectrum server,” in New Frontiers in Dynamic Spectrum Access Networks, 2005. DySPAN 2005. 2005 First IEEE International Symposium on, 2005, pp. 110–118.
[30] D. P. Bertsekas, Nonlinear Programming. Athena Scientific, Belmont, USA, 1999.
[31] M. Jaggi, “Revisiting frank-wolfe projection-free sparse convex optimization,” in Proceedings of the 30th International Conference on Machine Learning (ICML-13), 2013, pp. 427–435.
[32] S. Boyd and L. Vandenberghe, Convex Optimization. Cambridge University Press, New York, USA, 2004.
[33] 3GPP, “Further advancements for e-utra physical layer aspects (tr 36.814),” vol. v9.0.0, 2010.
[34] S. Deb, P. Monogioudis, J. Miernik, and J. Seymour, “Algorithms for enhanced inter-cell interference coordination (eicic) in LTE hetnets,” IEEE/ACM Trans. Netw., vol. 22, no. 1, pp. 137–150, 2014.
[35] V. Gajic, J. Huang, and B. Rimoldi, “Competition of wireless providers for atomic users: Equilibrium and social optimality,” in Communication, Control, and Computing, 2009. Allerton 2009. 47th Annual Allerton Conference on, 2009, pp. 1203–1210.