Approximation Algorithms for Online User Association in Multi-Tier Multi-Cell Mobile Networks

Weng Chon Ao and Konstantinos Psounis, Senior Member, IEEE, ACM

Abstract—The constantly growing wireless bandwidth demand is pushing wireless networks to multi-tier architectures consisting of a macrocell tier and a number of dense small cell deployment tiers. In such a multi-tier multi-cell environment, the classic problem of associating users to base stations becomes both more challenging and more critical to the overall network performance. Most previous analytical work is focused on offline/static user-cell association, where the users’ arrivals and their rates are assumed to be known in advance and thus has little practical relevance. On the other hand, practical online algorithms based on heuristics are often suboptimal and may not provide any performance guarantees.

In this paper, we propose online algorithms for the multi-tier multi-cell user association problem that have provable performance guarantees which improve previously known bounds by a sizable amount. The proposed algorithms are motivated by online combinatorial auctions, while capturing and leveraging the relative sparsity of choices in wireless networks as compared to auction setups. Our champion algorithm is a \( \frac{1}{1 \alpha} \) approximation algorithm, where \( \alpha \) is the maximum number of feasible associations for a user and is, in general, small due to path loss. Our analysis takes into account the state of the art wireless technologies such as massive and multiuser MIMO, and practical aspects of the system such as the fact that highly mobile users have a preference to connect to larger cell tiers to keep the signaling overhead low. In addition to establishing formal performance bounds, we also conduct simulations under realistic assumptions which establish the superiority of the proposed algorithm over existing approaches under real-world scenarios.

Index Terms—user association, load balancing, heterogeneous cellular networks, MIMO, mobility, online algorithm, randomized approximation algorithm.

I. INTRODUCTION

To support the tremendous growth of wireless data traffic fueled by popular applications like video streaming, enterprise networks consist of dense deployments of access points (APs), while a dense deployment of small cells (e.g., microcells and femtocells) under the coverage of macrocells has been proposed for future cellular networks in the upcoming 5G standard [1]. Such small cells could operate at a different frequency spectrum than macrocells (e.g., millimeter wave systems at 60 GHz [2]), and the performance of the overall cellular network can be sizably improved by this heterogeneous multi-tier architecture [3]. In addition, it is envisioned that antenna arrays will be deployed at cells to provide large spatial multiplexing gain with low-complexity linear precoding via a large number of antenna elements (massive MIMO) and/or via multiuser MIMO schemes, see, for example, [4]–[7].

In the context of such a multi-tier, multi-cell MIMO-enabled network, users typically have multiple choices when it comes to associating with a base station (BS). The fundamental problem of how to properly associate users with base stations so that the overall system performance is maximized is both more complex and more critical in such deployments. The association depends on many factors such as the quality of the received signal from the base stations at each user, the system load at the base stations, the user mobility, etc.

There is a large body of prior work in academia on the user-BS association problem. However, most prior work constitutes of offline analysis assuming full knowledge of the information of all users in advance (e.g., the number of users, the users’ arrivals, and the users’ rates), see, for example, [8]–[20] and references therein. This approach yields a static optimization framework which has limited practical relevance. What is more, in an effort to make such optimization problems more tractable, researchers have resorted to relaxation which leads to fractional solutions (where users are associated with multiple base stations and associate with each one of them for a fraction of time) and other non-practical ideas like solving the optimization problem from scratch every time there is a new user arrival, thus resulting in re-associating large numbers of pre-associated users.

On the other hand, practical online algorithms that are used in the industry are based on simple heuristics which waste precious system capacity and lead to suboptimal performance [12], while offering no performance guarantees. For example, by default, in todays cellular/WiFi networks users simply associate with the BS/AP from which they receive the strongest signal. And, some manufacturers of dense enterprise WiFi networks have recently attempted to impose some sort of load balancing by capping the maximum number of users an access point may associate with [21], while the LTE standard allows the introduction of a bias to offload users from macrocells to small cells when the latter are present, even when the signal from the macrocells is stronger. Note that in contrast to offline setups, in practical online algorithms the rates (or channel conditions) of a user from the base stations are revealed only when the user arrives. Then, the association decision is immediately and irrevocably made based only on the past user arrival profile.

In this paper, we propose novel online algorithms for the multi-tier user-BS association problem (the single tier user-
BS association problem is obviously a special case), which are both practical and provably near-optimal. The algorithms are motivated by online combinatorial auctions (bidders bid on objects) [22]–[24], where the base stations act as bidders and the users act as objects. By applying properties of wireless systems to the analysis of the online algorithms for combinatorial auctions, we are able to prove a performance guarantee which is close to the offline optimal. Specifically, we exploit the fact that a user can only receive and decode reference signals from a small number of nearby base stations due to path loss and interference. Therefore, the candidate set of feasible associations of a user is small, whereas in combinatorial auctions each bidder is in general assumed to have a positive valuation for every object. It turns out that by taking advantage of such “sparsity” together with introducing random decisions which favor “better” association candidates, our champion online algorithm achieves at least $\frac{1}{2}$ of the offline optimal, which, for typical values of $\alpha$, say 2 or 3, yields about 60 - 67% of the optimal performance guaranteed.

To the best of our knowledge, this is the tightest known bound achievable by online association algorithms, see Section II for more details.

The remainder of this paper is organized as follows. We present related work and highlight our contributions in Section II. Section III describes the system model where we consider both the massive and multiuser MIMO scenarios and formally state the offline association problem. In Section IV, we present our online multi-tier multi-cell user association algorithms. The performance analysis of the algorithms is presented in Section V. We discuss how the proposed algorithm can be applied in various practical scenarios of interest in Section VI. Section VII presents numerical and simulation results for a number of real-world scenarios. Last, section VIII concludes the paper.

II. PRIOR WORK AND CONTRIBUTION

We start with a discussion of prior work on offline user association schemes, followed by a discussion about prior work on online schemes.

A. Offline user association

Offline user association (also known as load balancing) has been well studied in the literature in the context of both WiFi networks and cellular networks, see, for example [8]–[20] and references therein. In general, under the offline setup, a static topology with users and base stations/access points is provided, and the association problem is formulated as an optimization problem. In the presence of new users arriving over time, the problem is solved from scratch each time a new user arrives.

In [8] the authors study the user-AP association problem ensuring a max-min fair bandwidth allocation. In [9] the authors perform joint AP channel selection and user association to minimize the user transmission delay. In [10], the authors associate users such that load balancing is achieved among APs. They achieve this by adjusting the power and thus the coverage of the APs. In [11], the authors propose a distributed user association policy that adapts to spatial traffic loads to achieve flow-level cell load balancing.

A recent overview of load balancing techniques in cellular networks can be found in [12]. In one of the works referred therein [13], the authors formulate the user-BS association problem as an integer programming problem. After relaxation of the integral constraints, the problem is reduced to a convex optimization problem, and dual algorithms are developed to iteratively solve for the optimal. While the relaxation leads to a plausible way to solve the optimization problem fast, it imposes unrealistic constraints as users end up associating with multiple base stations, spending a fraction of their time associated with each of them. In [14], the user-BS association problem is investigated in the context of massive MIMO wireless networks. Under the time scale over which the large-scale channel coefficients remain constant, the association problem is formulated as a network utility maximization problem that gives the fraction of time of a user associating to each base station. The problem is further extended to the case with base station cooperation in [15]. In [16], the authors study the user association problem in millimeter wave wireless networks operating at 60 GHz. In [17], [18] the multi-tier user-BS association problem is analyzed using stochastic geometry, and in [19] a game-theoretic model is proposed to associate users with different radio access technologies. Last, in [20], the authors propose approximation algorithms for the user-BS association problem to minimize the maximum load among all base stations.

All this prior work has limited relevance to practice since it studies the offline, static case where the complete setup is assumed to be known, it allows fractional associations to reduce the problem to a convex one, and it allows user re-associations to accommodate new user arrivals while maintaining high performance.

B. Online user association

Contrary to the offline algorithms, there is less related work on designing approximation algorithms for the online user-BS association problem. In [25], the authors propose a heuristic online algorithm for dynamic user association. In [26] the authors introduce a 1/8 approximation algorithm for online user-BS association to maximize the sum rate of the users under equal time sharing scheduling and equal power allocation, and, in [27] they introduce a 1/2 approximation algorithm to maximize the sum rate under a broadcast channel with receiver cooperation scenario and water-filling power allocation. Last, in [28] the authors derive an association algorithm aiming at minimizing the maximum load among all base stations. The performance bound of the proposed algorithm is proportional to the ratio of the minimum user rate over the maximum user rate, which for real world systems is more than 10, thus yielding an approximation bound which is a bit looser than 1/10.

In this paper, we consider the online multi-tier multi-cell user association with the objective of maximizing the sum utility of the users, which can be written as the sum of a number of “base station utility functions”. A base station
utility function is defined as the sum utility of its associated users. As a concrete example, we will analyze the logarithmic user utility (with respect to the data rate) with a bias of associating users with high mobility to tiers operating at low frequency with large cell coverage. Note that the logarithmic user utility captures the concept of proportional fairness [13].

In addition to the fact that proportional fairness is a good approximation of the operational point of today’s networks, under mild assumptions it also yields a monotone and submodular base station utility which renders the problem analytically tractable. The proposed online algorithm is proved to be a \( \frac{1}{2-a} \) approximation algorithm where the parameter \( a \) equals the maximum number of potential associations of a user. Note that the smaller the value of \( a \), the tighter the bound. (For \( a = 1 \) there is only one choice and there is no association decision that can be made for a user.) Due to path loss, signal degradation, interference in wireless medium, and the physical deployment of base stations, \( a \) is typically small, yielding a bound which is much tighter than the previous best known bound for an online association algorithm under realistic assumptions, and it is the tightest among all prior bounds.

III. SYSTEM MODEL

A. Network topology

Let \( \mathcal{U} = \{1, 2, \cdots, M\} \) be the set of users and the cardinality of \( \mathcal{U} \) be \( M \). Without loss of generality, we index the users according to their arrival to the system, i.e., user 1 arrives first and user \( M \) arrives last. Note that the proposed online algorithm does not need to know the total number of users \( M \). In other words, the performance guarantee holds for any instant of user arrival \( m, m = 1, 2, \cdots, M \). The users are just arriving online, and each user shall be associated upon arrival to one of the base stations.

We consider a multi-tier heterogeneous network with \( K \) tiers and we denote the set of tiers as \( \mathcal{K} = \{1, 2, \cdots, K\} \). We assume that there are \( N_k \) base stations (denoted as \( B_k = \{1, 2, \cdots, N_k\} \)) operating at tier \( k \in \mathcal{K} \). As a result, each base station is indexed by a tuple \((j, k)\), \( k \in \mathcal{K}, j \in B_k \).

The bandwidth of the spectrum band of the \( k \)-th tier is denoted as \( W_k \) and the spectrum bands of different tiers do not overlap. We consider a single carrier system where each base station in the \( k \)-th tier uses the whole spectrum band with bandwidth \( W_k \) for data transmission. (The analysis can be easily generalized to a multi-carrier system where the spectrum is divided into time-frequency slots (recourse blocks) as well as a multi-channel system with pre-allocated channels, see Section VII for a multi-channel scenario.) Since the base stations in the same tier share the same spectrum band, their transmissions will interfere with each other. Last, note that if different tiers are using the same spectrum band, e.g., as with today’s macro and small cells in cellular networks, the only change would be to replace the interference from a single tier with the sum of the interference from all tiers using the same spectrum band and the analysis would work the same way.

We consider the multi-tier cellular downlink user-BS association scenario depicted in Fig. 1. For each user \( i \in \mathcal{U} \), we define the set \( \mathcal{A}_i \) as the set of base stations that user \( i \) can potentially be associated with. Specifically, \( \mathcal{A}_i \) is the set of base stations from which the received SINR at user \( i \) is larger than some threshold \( \tau \) (which is chosen to ensure successful decoding of data messages), i.e.,

\[
\mathcal{A}_i \triangleq \{(j, k) : \text{SINR}_{i,j,k} \geq \tau, k \in \mathcal{K}, j \in B_k\}, \tag{1}
\]

where \( \text{SINR}_{i,j,k} \) is the received SINR at user \( i \) from base station \((j, k)\). The value of the received SINR (and thus the data rate) depends on the signaling scheme that we use. In the following two subsections, we consider two popular options. Specifically, we consider the signaling and the data rates in the massive MIMO regime and the multi-user (MU) MIMO full multiplexing gain regime, respectively.

B. Data rates in the massive MIMO regime

We assume that the system operates at the massive MIMO regime in which each base station is equipped with a large antenna array, while the users are assumed to be equipped with a single antenna. Let \( L_{j,k} \) denote the number of antennas at base station \((j, k)\) and \( \eta_{j,k} \) denote the number of users that base station \((j, k)\) can simultaneously service on any given time slot. In other words, \( \eta_{j,k} \) is the spatial multiplexing gain of base station \((j, k)\) and the ratio

\[
\mu_{j,k} \triangleq \frac{\eta_{j,k}}{L_{j,k}} \tag{2}
\]

is the corresponding spatial load [14]. We assume Time Division Duplex (TDD) operation with reciprocity-based channel state estimation. As a result, each base station antenna close to user \( i \) can estimate its downlink channel coefficient to user \( i \) from the uplink pilot transmitted by user \( i \), facilitating the training of large antenna arrays with training overhead proportional to \( \eta_{j,k} \).

In the massive MIMO regime, the value of \( \text{SINR}_{i,j,k} \) depends on the beamforming techniques that we use. We consider the following two commonly used schemes: Conjugate Beamforming (CB) and Zero-Forcing Beamforming (ZFBF). Under conjugate beamforming, the SINR can be expressed as

Fig. 1: A scenario of multi-tier user-BS association.
follows [4], [14]:
\[
\text{SINR}^{\text{CB}}_{i,j,k} = \frac{P_{j,k}g_{i,j,k}^2/\mu_{i,k}}{\nu W_k N_0 + \xi \sum_{l \in B_k} P_{l,k}g_{i,l,k} + \sum_{l \in B_k^{(i)}, l \neq j} P_{l,k}g_{i,l,k}/\mu_{i,k}},
\]
where \(P_{j,k}\) is the transmission power of base station \(j\) at tier \(k\), and \(g_{i,j,k}\) is the channel gain between user \(i\) and base station \(j\) at tier \(k\) that captures the effects of path loss and shadowing.

The effect of small-scale fading is modeled as Rayleigh fading coefficients. Note that the Rayleigh fading coefficients do not appear in Eq. (3) since in the massive MIMO regime, the effect of small-scale fading averages out over the antenna array (a fact commonly referred to as channel hardening). \(N_0\) is the noise power spectral density and \(\nu\) and \(\xi\) are normalization constants. \(\sum_{l \in B_k} P_{l,k}g_{i,l,k}\) is the interference received from base stations operating at the same tier. The set \(B_k^{(i)}\) denotes the set of base stations operating at tier \(k\) using the pilot signal \(q(i)\) that is also used by user \(i\). Hence, \(\sum_{l \in B_k^{(i)}, l \neq j} P_{l,k}g_{i,l,k}/\mu_{i,k}\) is the interference received from base stations operating at the same tier and using the same pilot signal as user \(i\) (commonly called pilot contamination).

Under ZFBF, the SINR is given by [14]:
\[
\text{SINR}^{\text{ZFBF}}_{i,j,k} = \left(1 - \mu_{i,k}\right) P_{j,k}g_{i,j,k}^2/\mu_{i,k} \times \left(\nu W_k N_0 + \xi \sum_{l \in B_k} P_{l,k}g_{i,l,k} + \sum_{l \in B_k^{(i)}, l \neq j} P_{l,k}g_{i,l,k}/\mu_{i,k}\right)^{-1},
\]
where \(1/\sigma^2\) is the SNR of the uplink pilot signal and the rest of the notation is like before. Note that the intra-cell interference \(\sigma^2 P_{j,k}g_{i,j,k}\) is zero when the uplink SNR \(1/\sigma^2 \rightarrow \infty\). The other two terms correspond to inter-cell interference and pilot contamination, respectively.

The data rate (bits/s) between user \(i \in \mathcal{U}\) and base station \((j, k) \in \mathcal{A}_i\) is given by
\[
c_i,j,k = W_k \log \left(1 + \text{SINR}^{\text{CB}/\text{ZFBF}}_{i,j,k}\right), \quad i \in \mathcal{U}, \ (j, k) \in \mathcal{A}_i,
\]
where Shannon’s formula is used which can be extended to accommodate real world features like modulation and coding tables, see, for example, [29].

C. Data rates in the MU-MIMO full multiplexing gain regime
We assume that base station \((j, k)\) has \(L_{j,k}\) antennas and the users are equipped with a single antenna. We assume that ZFBF is used for MU-MIMO beamforming and we are able to use all degrees of freedom. In other words, the base station \((j, k)\) can provide a full multiplexing gain of order \(\eta_{j,k} = L_{j,k}\) to support \(L_{j,k}\) simultaneous data transmissions/streams to its associated users. Similar to [30], [31], under equal power allocation on each data stream and by using random matrix theory, we obtain the following deterministic approximation for the SINR:
\[
\text{SINR}^{\text{MU-MIMO}}_{i,j,k} = \frac{P_{j,k}g_{i,j,k}}{W_k N_0 + \sum_{l \in B_k, l \neq j} P_{l,k}g_{i,l,k}}.
\]
Then, the data rate (bits/s) between user \(i\) and base station \((j, k)\) is given by
\[
c_{i,j,k} = W_k \log \left(1 + \text{SINR}^{\text{MU-MIMO}}_{i,j,k}\right), \quad i \in \mathcal{U}, \ (j, k) \in \mathcal{A}_i.
\]

D. User scheduling
Let the association variable be \(x_{i,j,k}\), where \(x_{i,j,k} = 1\) if user \(i\) is associated with base station \((j, k) \in \mathcal{A}_i\) and \(x_{i,j,k} = 0\) otherwise. The actual data rate that user \(i\) will receive, which is denoted as \(r_{i,j,k}\), depends on the user scheduling mechanism. We assume that when a base station is associated with multiple users and the number of the associated users is larger than the spatial multiplexing gain \(\eta_{j,k}\), equal time-sharing is used to schedule the users. (This is not only what happens in most real-world systems, but also it is the optimal schedule under our scenario, see Section V-B for further details and a proof.) Specifically, we have
\[
r_{i,j,k} = c_{i,j,k}, \quad \text{if} \quad \sum_{l \in \mathcal{U}} x_{l,j,k} \leq \eta_{j,k},
\]
and
\[
r_{i,j,k} = \frac{\eta_{j,k} c_{i,j,k}}{\sum_{l \in \mathcal{U}} x_{l,j,k}}, \quad \text{if} \quad \sum_{l \in \mathcal{U}} x_{l,j,k} > \eta_{j,k}.
\]
Furthermore, the utility function of user \(i\) is denoted as \(U_i(r_{i,j,k}, v_i, z_{j,k})\), which is a function of the actual data rate \(r_{i,j,k}\), the speed of the user \(v_i\), and the coverage of base station \((j, k)\), \(z_{j,k}\). The multi-tier user-BS association problem is to find the association such that the sum utility of the users is maximized.

E. Offline user-BS association
We first consider the offline multi-tier user-BS association problem (denoted as \(Q\)):
\[
\text{Q} : \text{maximize} \sum_{i \in \mathcal{U}} \sum_{(j, k) \in \mathcal{A}_i} x_{i,j,k} U_i(r_{i,j,k}, v_i, z_{j,k})
\]
subject to
\[
\sum_{(j, k) \in \mathcal{A}_i} x_{i,j,k} = 1, \quad i \in \mathcal{U},
\]
\[
x_{i,j,k} \in \{0, 1\}, \quad i \in \mathcal{U}, \ (j, k) \in \mathcal{A}_i,
\]
\[
r_{i,j,k} = \begin{cases} c_{i,j,k} & \text{if} \sum_{l \in \mathcal{U}} x_{l,j,k} \leq \eta_{j,k} \\ \eta_{j,k} c_{i,j,k}/\sum_{l \in \mathcal{U}} x_{l,j,k} & \text{if} \sum_{l \in \mathcal{U}} x_{l,j,k} > \eta_{j,k}, \end{cases}
\]
where the first constraint ensures that a user can only be associated with a single base station. We denote the value of the offline optimal as \(\text{OPT}(Q)\).
Since the above problem \( Q \) is an integer program, it is hard to find the optimal solution in general. To facilitate the comparison between the performance of the offline algorithm and online algorithms, we derive an upper bound of \( OPT(Q) \) by considering the relaxation of problem \( Q \) (denoted as \( \tilde{Q} \)) that allows fractional association:

\[
\tilde{Q} : \text{maximize} \sum_{i \in U} \sum_{(j,k) \in A_i} x_{i,j,k} c_{i,j,k}, v_i, z_j \]

subject to

\[
\begin{align*}
& x_{i,j,k} \leq 1, \quad i \in U \\
& x_{i,j,k} \geq 0, \quad i \in U, \quad (j,k) \in A_i \\
& \sum_{i \in U} x_{i,j,k} \leq \eta_{j,k}, \quad k \in K, \quad j \in B_k.
\end{align*}
\]

We denote the optimal value of the problem \( \tilde{Q} \) as \( OPT(\tilde{Q}) \), and clearly we have \( OPT(Q) \leq OPT(\tilde{Q}) \). In addition, it can be observed that when the utility function is concave in the data rate, the problem \( \tilde{Q} \) becomes a convex optimization problem and can be solved efficiently. Last, note that since the offline algorithm recomputes the optimal association every time there is a new arrival or departure, it is clearly optimal in the long run.

**F. Special cases**

When \( \eta_{j,k} = 1 \), the above problem \( Q \) can be simplified as

\[
\begin{align*}
\text{maximize} & \sum_{i \in U} \sum_{(j,k) \in A_i} x_{i,j,k} c_{i,j,k}, v_i, z_j \\
\text{subject to} & \sum_{(j,k) \in A_i} x_{i,j,k} = 1, \quad i \in U \\
& x_{i,j,k} \leq 1, \quad i \in U, \quad (j,k) \in A_i.
\end{align*}
\]

In the massive MIMO regime, the special case with \( \eta_{j,k} = 1 \) (and thus \( \rho_{j,k} = \frac{1}{L_{j,k}} \)) corresponds to having an array gain of order \( L_{j,k} \) for the desired signal but not having any multiplexing gain. In the MU-MIMO full multiplexing gain regime, the special case with \( \eta_{j,k} = L_{j,k} = 1 \) corresponds to a point-to-point single-input single-output (SISO) channel, which is the specific scenario we consider in [32] (without using a bias to associate users with high mobility to large cells, like we do in this work).

**IV. ONLINE ASSOCIATION ALGORITHMS**

In the following, we consider three online algorithms for the multi-tier user-BS association, where the users arrive online (user 1 arrives first and user \( M \) arrives last) and the association decision is immediately and irrevocably made upon each user’s arrival. The first online algorithm is user-centric in that the user makes a decision based on its own performance. The second, which is the algorithm we advocate, is cell-centric in the sense that the association decision strives to maximize the performance of cells, and the third is a deterministic, somewhat simplified version of the second.

**A. User-centric online algorithm**

In the user-centric algorithm (Algorithm 1), when a user arrives, the user is associated with the base station that maximizes the user’s own utility. The variable \( s_{j,k} \) updates the number of users associated with base station \( j \) at tier \( k \). Note that at the end of the algorithm, we have \( \sum_{k \in K} \sum_{j \in B_k} s_{j,k} = M \). In practice, when a user arrives, the user can obtain information of the system load \( s_{j,k} \), the cell range \( z_{j,k} \), and the available degrees of freedom \( \eta_{j,k} \) by base station broadcast and the data rate \( c_{i,j,k} \) by training, sensing and estimation, see, for example, [21]. We denote the resulting sum utility of the users under the user-centric online algorithm as \( ALG_1(Q) \).

**Algorithm 1 User-centric online algorithm**

1. Initialize \( s_{j,k} \leftarrow 0 \), \( k \in K, \quad j \in B_k \);
2. for \( i = 1, \ldots, M \) do
3. Associate user \( i \) with base station \( j^* \) at tier \( k^* \), where
   \[
   (j^*, k^*) = \arg\max_{(j,k) \in A_i} U_i \left( r_{i,j,k}, v_i, z_j \right),
   \]
   where \( r_{i,j,k} = c_{i,j,k} \) if \( s_{j,k} + 1 \leq \eta_{j,k} \), otherwise \( r_{i,j,k} = \eta_{j,k} c_{i,j,k} / (s_{j,k} + 1) \);
4. \( s_{j^*, k^*} \leftarrow s_{j^*, k^*} + 1 \);
5. end for

**B. Cell-centric randomized online algorithm**

To facilitate analysis, let us first introduce the concept of the utility of a base station. The utility of the base station \( j \) at tier \( k \) (denoted as \( V_{j,k} \)) is defined as the sum utility of its associated users. The domain of \( V_{j,k} \) (denoted as \( A_{j,k} \)) is the set of users that the base station \( j \) at tier \( k \) can be associated with, i.e.,

\[
A_{j,k} = \{ i \in U : (j,k) \in A_i \}.
\]

We have

\[
V_{j,k}(S) = \sum_{i \in S} U_i \left( r_{i,j,k}, v_i, z_j \right), \quad k \in K, \quad j \in B_k, \quad S \subset A_{j,k},
\]

\[
r_{i,j,k} = c_{i,j,k} 1_{|S| \leq \eta_{j,k}} + \frac{\eta_{j,k} c_{i,j,k}}{|S|} 1_{|S| > \eta_{j,k}},
\]

where \( S \) denotes the set of users that base station \( (j,k) \) is associated with, \( |S| \) is the cardinality of \( S \), and \( 1_{\{\cdot\}} \) is the indicator function. In addition, we let \( V_{j,k}(\emptyset) = 0 \). We further define the marginal utility of the base station \( j \) at tier \( k \) for associating with a “new” user \( i \) given the set of “previously” associated users \( S \) as

\[
V_{j,k}(i|S) = V_{j,k}(S \cup \{i\}) - V_{j,k}(S), \quad i \in A_{j,k}, \quad S \subset A_{j,k}, \quad i \notin S.
\]

In the cell-centric randomized algorithm (Algorithm 2), when a user arrives, the user is associated with a base station in a probabilistic manner. Specifically, the probability of associating a user with a base station is proportional to the base station’s marginal utility (of including that user). In this sense, a user will most likely be associated with the base station with the highest marginal utility. The variable \( S_{j,k} \) updates the set of users that base station \( (j,k) \) is associated with. At the end of the algorithm, the sets \( S_{j,k}, \quad k \in K, \quad j \in B_k \) form
Algorithm 2 Cell-centric randomized online algorithm

1: Initialize $S_{j,k} \leftarrow \emptyset$, $k \in K$, $j \in B_k$;
2: for $i = 1, \ldots, M$ do
3: Associate user $i$ with base station $j$ at tier $k$ with probability
4: 
5: 
6: Let the selected base station be $(j^*, k^*)$;
7: $S_{j^*, k^*} \leftarrow S_{j^*, k^*} \cup \{i\}$;
8: end for

Algorithm 3 Cell-centric deterministic online algorithm

1: Initialize $S_{j,k} \leftarrow \emptyset$, $k \in K$, $j \in B_k$;
2: for $i = 1, \ldots, M$ do
3: Associate user $i$ with base station $j^*$ at tier $k^*$, where
4: $(j^*, k^*) = \arg \max_{(j,k) \in A_i} V_{j,k}(i|S_{j,k})$;
5: $S_{j^*, k^*} \leftarrow S_{j^*, k^*} \cup \{i\}$;
6: end for

randomized version (Algorithm 2), the deterministic version (Algorithm 3) is easier to implement. However, it will be shown that the deterministic version has a worse performance guarantee than the randomized one. We denote the resulting sum utility of the users under the cell-centric deterministic online algorithm as $ALG_2(Q)$.

V. PERFORMANCE ANALYSIS

In this section we establish the performance bound for the two cell-centric online algorithms using the theory of online combinatorial auctions. The main notation is summarized in Table I. To apply results from online combinatorial auctions, we first need to prove that under mild assumptions, the specific base station utility function for our application, namely $V_{j,k}(\cdot)$ in Eq. (14), is submodular and monotone. As a concrete example, we consider the following user utility

$$U_i(r_{i,j,k}, v_i, z_{j,k}) = \log(r_{i,j,k}) + f(v_i, z_{j,k}).$$

The user utility consists of two parts: the logarithmic user utility ($\log(r_{i,j,k})$) with respect to the data rate, which is commonly used in wireless networks to provide proportional fairness among users [13], and the bias $f(v_i, z_{j,k})$ of associating users with high mobility to tiers operating at low frequency with large cell coverage, which depends on the user mobility ($v_i$) and the cell coverage ($z_{j,k}$). The bias function $f(x, y)$ is defined on the domain $x \geq 0, y \geq 0$, satisfying

1. $f(x, y) \geq 0$;
2. For any given $x$, $f(x, y_1) \geq f(x, y_2), \forall y_1 \geq y_2.$

The bias function $f(x, y)$ defined above implies that a user with a high speed/mobility prefers to be associated with a cell with a large coverage, see Fig. 2 for an example. This makes sense from a practical point of view as the higher the speed the more frequent the handoffs from one cell to another, thus larger cells keep the signaling overhead at reasonable levels.

Under the user utility of Eq. (18), the base station utility function becomes

$$V_{j,k}(S) = \begin{cases} \sum_{i \in S} \log (r_{i,j,k}) + f(v_i, z_{j,k}) & \text{if } |S| \leq \eta_{j,k} \\ \sum_{i \in S} \log \left( \frac{\eta_{j,k} c_{i,j,k}}{|S|} \right) + f(v_i, z_{j,k}) & \text{if } |S| > \eta_{j,k}, \end{cases}$$

$k \in K, j \in B_k, S \subset A_{j,k}$.

The coverage of a cell depends on a number of factors such as the power, the carrier frequency, and the elevation of the tower.

<table>
<thead>
<tr>
<th>TABLE I: Main notation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data rate that user $i$ can get if associated with BS $(j, k)$</td>
</tr>
<tr>
<td>Maximum data rate (no time sharing) that user $i$ can get if associated with BS $(j, k)$</td>
</tr>
<tr>
<td>User speed</td>
</tr>
<tr>
<td>BS coverage</td>
</tr>
<tr>
<td>Spatial multiplexing gain</td>
</tr>
<tr>
<td>User utility</td>
</tr>
<tr>
<td>The bias function</td>
</tr>
<tr>
<td>Base station utility</td>
</tr>
<tr>
<td>Marginal base station utility</td>
</tr>
<tr>
<td>The set of users that BS $(j, k)$ can be associated with</td>
</tr>
<tr>
<td>The set of BSs that user $i$ can be associated with</td>
</tr>
</tbody>
</table>

Fig. 2: Consider three cells with cell range $y_1 > y_2 > y_3$. While for slow-moving users it is indifferent to which cell they will associate with when it comes to signaling overhead, for high-speed users the larger the cell range the better.
In addition, the marginal base station utility function can be derived as follows: If $|S| \leq \eta_{j,k} - 1$, we have

$$V_{j,k}(i|S) = \log(c_{i,j,k}) + f(v_i, z_{j,k});$$

(21)

If $|S| \geq \eta_{j,k}$, we have

$$V_{j,k}(i|S) = V_{j,k}(S \cup \{i\}) - V_{j,k}(S)$$

$$= \sum_{t \in S \cup \{i\}} \left( \log \left( \frac{\eta_{j,k}c_{t,j,k}}{|S|+1} \right) + f(v_i, z_{j,k}) \right)$$

$$- \sum_{t \in S} \left( \log \left( \frac{\eta_{j,k}c_{t,j,k}}{|S|} \right) + f(v_i, z_{j,k}) \right)$$

$$= \log(\eta_{j,k}c_{i,j,k}) + f(v_i, z_{j,k}) + |S| \log |S|$$

$$- (|S| + 1) \log(|S| + 1).$$

(22)

**Definition 1.** The base station utility function $V_{j,k}(\cdot)$ is submodular if $V_{j,k}(i|S) \geq V_{j,k}(i|T)$ for all $i \in A_{j,k}$, $S \subset T \subset A_{j,k}$, $i \notin T$.

**Definition 2.** The base station utility function $V_{j,k}(\cdot)$ is monotone if $V_{j,k}(i|S) \geq 0$ for all $i \in A_{j,k}$, $S \subset A_{j,k}$, $i \notin S$.

**Lemma 1.** The base station utility function $V_{j,k}(\cdot)$ in Eq. (20) is submodular.

**Proof.** Let $i \in A_{j,k}$, $S \subset T \subset A_{j,k}$, $i \notin T$ be given. There are three cases.

In the first case, we consider $|T| \leq \eta_{j,k} - 1$. Clearly, we have $V_{j,k}(i|S) = V_{j,k}(i|T)$.

In the second case, we consider $|S| \leq \eta_{j,k} - 1$ and $|T| \geq \eta_{j,k}$. We have

$$V_{j,k}(i|S) - V_{j,k}(i|T)$$

$$= - \log(\eta_{j,k}) - |T| \log |T| + (|T| + 1) \log(|T| + 1)$$

$$= \log \left( \frac{|T| + 1}{|T|^{|T|+1}} \right) \eta_{j,k}$$

$$\geq \log \left( \frac{|T| + 1}{|T|^{|T|+1}} \right) > 0.$$ 

(23)

In the third case, we consider $|S| \geq \eta_{j,k}$. To check that $V_{j,k}(i|S) \geq V_{j,k}(i|T)$, it is equivalent to show that

$$|S| \log |S| - (|S| + 1) \log(|S| + 1) \geq |T| \log |T| - (|T| + 1) \log(|T| + 1),$$

which in turn is equivalent to show that the function $h(x) \equiv x \log x - (x + 1) \log(x + 1), x > 0$, is decreasing. Indeed, we have

$$h'(x) = \log x - \log(x + 1)$$

$$= - \log \left( \frac{x + 1}{x} \right) < 0, \ \forall x > 0,$$

(24)

which implies that $h(x)$ is decreasing.

As a result, in all cases we have $V_{j,k}(i|S) \geq V_{j,k}(i|T)$.

We conclude that $V_{j,k}(\cdot)$ is submodular.

**Lemma 2.** If $c_{i,j,k} \geq \max \left\{ \frac{|A_{j,k}|}{\eta_{j,k}} e^{1-f(v_i, z_{j,k})}, 1 \right\}$ bits/s, $\forall i \in A_{j,k}$, then $V_{j,k}(\cdot)$ in Eq. (20) is monotone.

**Note:** For the monotonicity to hold we need $c_{i,j,k} \geq \max \left\{ \frac{|A_{j,k}|}{\eta_{j,k}} e^{1-f(v_i, z_{j,k})}, 1 \right\}$ bits/s, $\forall i \in A_{j,k}$, i.e., we need the data rate (measured in bits/s) between base station $(j, k)$ and user $i \in A_{j,k}$ to be larger than $\frac{e^{1-f(v_i, z_{j,k})}}{\eta_{j,k}} \times$ the number of users that base station $(j, k)$ can be associated with, which is trivially satisfied for any real world scenario.

**Proof.** Let $i \in A_{j,k}$, $S \subset A_{j,k}$, $i \notin S$ be given. From Eq. (22), we have

$$V_{j,k}(i|S) = \log(\eta_{j,k}c_{i,j,k}) + f(v_i, z_{j,k})$$

$$+ |S| \log |S| - (|S| + 1) \log(|S| + 1)$$

$$\geq \log(\eta_{j,k}c_{i,j,k}) + f(v_i, z_{j,k})$$

$$+ (|A_{j,k}| - 1) \log(|A_{j,k}| - 1) - |A_{j,k}| \log|A_{j,k}|$$

$$= \log(\eta_{j,k}c_{i,j,k}) + f(v_i, z_{j,k}) - \log \left( \frac{|A_{j,k}|}{|A_{j,k}| - 1} \right)$$

$$\geq \log(\eta_{j,k}c_{i,j,k}) + f(v_i, z_{j,k}) - \log |A_{j,k}|,$$

(25)

where (a) holds since the function $h(x) \equiv x \log x - (x + 1) \log(x + 1), x > 0$ is decreasing and achieves its minimum when $|S| = |A_{j,k}|-1$.

Therefore, if $c_{i,j,k} \geq \max \left\{ \frac{|A_{j,k}|}{\eta_{j,k}} e^{1-f(v_i, z_{j,k})}, 1 \right\}$ bits/s, $\forall i \in A_{j,k}$, we have $V_{j,k}(i|S) \geq 0$ and thus $V_{j,k}(\cdot)$ is monotone.

**A. Performance bounds**

We first derive the performance bound of the cell-centric randomized online algorithm.

**Theorem 1.** Under the submodularity and monotonicity of the base station utility function $V_{j,k}(\cdot)$, we have $E[ALG_{2}(Q)] \geq \frac{1}{2-a} OPT(Q)$, where $a \equiv \max_{i \in Q} |A_i|$.

**Proof.** After establishing the submodularity and monotonicity of the base station utility function $V_{j,k}(\cdot)$, one may apply some somewhat recent results from online combinatorial auctions, see [23], to get a lower bound equal to $\frac{1}{2-a} OPT(Q)$, where $N = \sum_{k=1}^{K} N_k$ is the total number of base stations in $K$ tiers (which could be very large). We further tighten this bound by exploiting the “sparsity” of feasible associations of a user in a heterogeneous wireless cellular system, and show that $E[ALG_{2}(Q)] \geq \frac{1}{2-a} OPT(Q)$, where $a = \max_{i \in \mathbb{Q}} |A_i|$ is the maximum number of potential associations of a user (see Eq. (1)). Clearly, since the bound deteriorates as $N$ and $a$ increase, smaller values of $a$ yield tighter bounds (we assume $a > 1$ since if $a=1$ there is no decision to be made).

We prove the performance bound by induction on the number of users $M$. Let $Q$ be the original problem of associating $M$ users to base stations. For each $(j, k) \in A_1$, we define $Q_{j,k}$ as the subproblem of associating the remaining users $2, \ldots, M$ to the base stations, where the base station utility function $V_{j,k}(\cdot)$ is replaced by $V_{j,k}(\{\{1\}\})$ (which is also a monotone
submodular function). From the cell-centric randomized online algorithm, we have
\[
E[ALG_2(Q)] = \sum_{(j,k)\in A_1} q_{j,k} \{ E[ALG_2(Q_{j,k})] + V_{j,k}(\{1\}) \},
\]
where
\[
q_{j,k} = \frac{V_{j,k}(\{1\})^{|A_1|-1}}{\sum_{(j',k')\in A_1} V_{j',k'}(\{1\})^{|A_1|-1}}, \quad (j,k) \in A_1.
\]
Let \( S = \{ S_{j,k}, k \in K, j \in B_k \} \) be the optimal offline association profile for the original problem \( Q \) and let us assume that user 1 is removed. Let us denote the value (the achieved sum user utility) of the subproblem \( Q_{j,k} \) under the association profile \( S' \) which is the same as \( S \) except that user 1 is removed. Let us denote the value (the achieved sum user utility) of the subproblem \( Q_{j,k} \) under the association profile \( S' \) as \( Val(Q_{j,k}) \). Obvioulsy, we have \( Val(Q_{j,k}) \leq OPT(Q_{j,k}) \) By the submodularity and monotonicity of \( V_{j,k}(\cdot) \), for all \((j,k)\in A_1, (j,k)\neq(j',k')\), we have \( OPT(Q) - Val(Q_{j,k}) \leq V_{j,k}(\{1\}) + V_{j,k}(\{1\}) \), where \( V_{j,k}(\{1\}) \) is the maximum “loss” due to the fact that the subproblem \( Q_{j,k} \) does not have user 1 associated with base station \((j,k)\), and \( V_{j,k}(\{1\}) \) is the maximum “loss” due to the fact that the subproblem \( Q_{j,k} \) uses the utility function \( V_{j,k}(\cdot|\{1\}) \) (instead of \( V_{j,k}(\cdot) \) in the original problem \( Q \)). For the case \((j,k)\) is the maximum “loss” due to the fact that the subproblem \( Q_{j,k} \) does not have user 1 associated with base station \((j,k)\), and \( V_{j,k}(\{1\}) \) is the maximum “loss” due to the fact that the subproblem \( Q_{j,k} \) uses the utility function \( V_{j,k}(\cdot|\{1\}) \) (instead of \( V_{j,k}(\cdot) \) in the original problem \( Q \)). As a result, we have
\[
OPT(Q) - \sum_{(j,k)\in A_1} q_{j,k} OPT(Q_{j,k})
\]
\[
\leq OPT(Q) - \sum_{(j,k)\in A_1} q_{j,k} Val(Q_{j,k})
\]
\[
\leq \sum_{(j,k)\in A_1} q_{j,k} [V_{j,k}(\{1\}) + V_{j,k}(\{1\})] + q_{j,k} V_{j,k}(\{1\})
\]
\[
= 1 + \frac{V_{j,k}(\{1\})^{|A_1|-1}}{\sum_{(j',k')\in A_1} V_{j',k'}(\{1\})^{|A_1|-1}} \cdot \frac{1}{1 - \frac{1}{a}}, \quad (28)
\]
where (a) follows by the AM-GM inequality (see Appendix).

Therefore, we have
\[
OPT(Q) \leq \sum_{(j,k)\in A_1} q_{j,k} OPT(Q_{j,k})
\]
\[
+ \left(2 - \frac{1}{a}\right) \sum_{(j,k)\in A_1} q_{j,k} V_{j,k}(\{1\})
\]
\[
\leq \sum_{(j,k)\in A_1} q_{j,k} \left(2 - \frac{1}{a}\right) [E[ALG_2(Q_{j,k})] + V_{j,k}(\{1\})]
\]
\[
\leq \left(2 - \frac{1}{a}\right) E[ALG_2(Q)], \quad (29)
\]
where (a) follows from Eq. (28), (b) follows by induction, and (c) follows from Eq. (26).

Now, we proceed to derive the performance bound of the cell-centric deterministic online algorithm.

**Theorem 2.** Under the submodularity and monotonicity of \( V_{j,k}(\cdot) \), we have \( ALG_3(Q) \geq \frac{1}{2} OPT(Q) \).

**Proof.** The submodularity and monotonicity of \( V_{j,k}(\cdot) \) are respectively shown in Lemma 1 and Lemma 2. Then, the \( 1/2 \)-performance guarantee follows by the analysis above and a result in online combinatorial auctions, see Theorem 11 in [22].

**Remark:** It is evident from the proofs of Theorems 1 and 2 that the specific form of the utility function does not play a role in the proof as long as the function is submodular and monotone. Thus, the above performance bounds hold for a generic submodular and monotone base station utility function \( V_{j,k}(\cdot) \) and not just for the logarithmic user utility function with bias which has been introduced in Eq. (20) as a concrete example. Also, recall that for this particular utility function to be monotone we need \( c_{i,j,k} \geq \max \{ |A_{j,k}|a^{-1}f(v_{i,j,k}), 1 \} \) bits/s, \( \forall i \in A_{j,k} \) which is trivially satisfied in practice.

**B. Rationale of equal time allocation**

When the number of the associated users \( |S| \) at base station \((j,k)\) is less than or equal to its spatial multiplexing gain \( \eta_{j,k} \), each associated user can be active for the whole duration without the need of time sharing. However, when \( |S| > \eta_{j,k} \), some kind of time sharing is needed.

In the above analysis, we assume that equal time sharing is used to schedule transmissions for users associated with the same base station when \( |S| > \eta_{j,k} \) (see Eq. (20)). To motivate this assumption, we generalize equal time sharing to a more flexible resource allocation scheme, in which different users are allowed to have different time portions for data transmissions, and show that under a logarithmic user utility with bias, equal time sharing is optimal.

For any base station \((j,k)\), \( k \in K, j \in B_k \), let \( S \subset A_{j,k} \) be the set of users associated with it and assume that \( |S| > \eta_{j,k} \). Let us define the time sharing variables \( w_{i,j,k}, i \in S \) where \( \sum_{i \in S} w_{i,j,k} = \eta_{j,k} \) and \( 0 \leq w_{i,j,k} \leq 1, \ i \in S \). The time sharing variables are optimized such that the sum utility of the users in \( S \) is maximized. In other words, when \( |S| > \eta_{j,k} \), the base station utility function is generalized from Eq. (20) to
\[
V_{j,k}(S) = \maximize_{w_{i,j,k}} \sum_{i \in S} w_{i,j,k} \log (w_{i,j,k} c_{i,j,k}) + f(v_i, z_j, k)
\]
subject to \( \sum_{i \in S} w_{i,j,k} = \eta_{j,k} \)
\[
0 \leq w_{i,j,k} \leq 1, \ i \in S. \quad (30)
\]
Let us define the Lagrange function
\[
L(w_{i,j,k}, \lambda) = \sum_{i \in S} \log (w_{i,j,k} c_{i,j,k}) + f(v_i, z_j, k)
\]
\[ - \lambda \left( \sum_{i \in S} w_{i,j,k} - \eta_{j,k} \right), \quad (31)
\]
where $\lambda$ is the Lagrange multiplier. By taking the derivative of $L(w_{i,j,k}, \lambda)$ with respect to $w_{i,j,k}$ and setting the result to zero, we have
\[
\frac{\partial L}{\partial w_{i,j,k}} = \frac{1}{w_{i,j,k}} - \lambda = 0 \Rightarrow w_{i,j,k} = \frac{1}{\lambda}.
\] (32)

Therefore, we have
\[
\sum_{i \in S} w_{i,j,k} = \frac{|S|}{\lambda} \Rightarrow \lambda = \frac{|S|}{\sum_{i \in S} w_{i,j,k}} \Rightarrow w_{i,j,k} = \frac{\eta_{j,k}}{|S|}.
\] (33)

We can see that $0 \leq \frac{\eta_{j,k}}{|S|} \leq 1$, so the optimal time sharing variables are indeed $w_{i,j,k} = \frac{\eta_{j,k}}{|S|}$, $i \in S$, showing that equal time allocation is optimal.

VI. EXTENSIONS

In the following, we comment on how the proposed cell-centric randomized online algorithm can be applied into scenarios with user heterogeneity, departing users, and base station cooperation.

A. Heterogeneous users and user priority

Heterogeneous users refer to users that subscribe at different services. For example, some users are allowed to connect to all $K$ tiers while others are restricted to connect to one tier. Similarly, users can be divided into different classes with different priorities. For example, primary users with high priority are allowed to access all base stations while secondary users with low priority are not [33]. Both heterogeneous users and user priority can be incorporated into the analysis by restricting the set of tiers and/or base stations which user $i$ may be associated with in Eq. (1), while the rest of the analysis remains unchanged.

B. Departing users

The performance bound on the cell-centric randomized algorithm holds as users arrive online. However, when users leave the system, the performance bound may no longer hold. A simple way to guarantee the bound when a user leaves is to backtrack to the association profile just before this user’s arrival, and consider re-associating users which arrived after this user. Specifically, suppose that there are $M$ users in the system, where as previously discussed user 1 arrived first, user $M$ arrived last, and they were associated with base stations by using Algorithm 2. Suppose now user $m$ leaves the system. We first backtrack to the association profile just before user $m$’s arrival (i.e., the association profile $S^{m-1} = \{S_{j,k}^{m-1}, k \in K, j \in B_k\}$) generated at the $m-1$th iteration of Algorithm 2 and then re-associate users $m+1$ to $M$. Clearly, this may result in a number of re-associations, which is not practical. In Section VII, we show that Algorithm 2 in the presence of user departures performs very close to the offline optimal, thus in practice there is no need to backtrack.

C. Base station cooperation

A dense deployment of small cells may yield even higher throughput when multiple neighboring base stations can cooperate with each other, an architecture often referred to as Coordinated Multi-Point (CoMP) [34]–[38], to form a cluster and coordinate their data transmissions such that they aggregate constructively. In a typical scenario of a two-tier heterogeneous network consisting of macro-BSs and femto-BSs, one may have tens or hundreds of femto-BSs inside a macrocell and hundreds or thousands of users. Thus, femto-BSs could be grouped into clusters of nearby femto-BSs which can concurrently serve a number of users. For example, one may have one such cluster per floor on a large building or one cluster per building. Along these lines, assuming that lower-power base stations form cooperation clusters, the user-BS association problem can be generalized to a user-cluster association problem (note that it is possible that a cluster is just a single base station).

While it is beyond the scope of this paper to investigate clustering algorithms, we wish to extend our association algorithms to make them applicable to the CoMP setup. Let the set of base stations at tier $k$, $B_k$, be partitioned into $G_k$ clusters $C_1, C_2, \ldots, C_{G_k}$, where $C_m \cap C_n = \emptyset$ and $\bigcup_{m=1}^{G_k} C_m = B_k$. We index the $m$-th cluster at tier $k$ by the tuple $(C_m, k)$. Suppose that distributed MU-MIMO ZFBF is used by the base stations in a cooperation cluster, say, $(C_m, k)$, to provide $\sum_{j \in C_m} \eta_{j,k}$ degrees of freedom for spatial multiplexing [38]. Similar to [31] and using the same asymptotic regime as the one we used to derive Eq. (6), the SINR at user $i$ from cluster $(C_m, k)$ becomes
\[
\text{SINR}_{i,C_m,k}^{\text{CoMP}} = \frac{g_{i,C_m,k}}{W_k N_0 + \sum_{l \in B_k \setminus C_m} P_l g_{l,i,l,k}},
\] (34)

where
\[
g_{i,C_m,k} = \frac{\left(\sum_{j \in C_m} P_{j,k}\right) \left(\sum_{j \in C_m} g_{j,i,k} / |C_m|\right)}{\sum_{j \in C_m} \eta_{j,k}}.
\] (35)

As a result, the data rate becomes $c_{i,C_m,k} = W_k \log (1 + \text{SINR}_{i,C_m,k}^{\text{CoMP}})$. In addition, the “cluster” utility function can be stated as follows (see Eq. (20)):
\[
V_{C_m,k}(S) = \sum_{i \in S} \log (c_{i,C_m,k}) + f(v_i, z_{C_m,k});
\] (36)

If $|S| \leq \sum_{j \in C_m} \eta_{j,k}$, we have
\[
V_{C_m,k}(S) = \sum_{i \in S} \log \left(\frac{c_{i,C_m,k} \sum_{j \in C_m} \eta_{j,k}}{|S|}\right) + f(v_i, z_{C_m,k}),
\] (37)

where $z_{C_m,k}$ is the coverage of the cluster $(C_m, k)$. By a straightforward extension of Lemmas 1 and 2 it is easy to show that this utility function is monotone and submodular as well, and the rest of the analysis goes through as before. Note that under this CoMP setup, the “sparsity” parameter $\alpha$ is the maximum number of potential cluster associations for a user, and, since this is naturally smaller than the number of potential base station associations, the cell-centric randomized online
algorithm has, in practice, an even tighter bound \( \frac{1}{2^{\frac{1}{n}}} \) with respect to the optimal than it had before.

VII. SIMULATION RESULTS

A. Two-tier heterogeneous cellular network in massive MIMO scenario

We consider a two-tier heterogeneous cellular network consisting of macro-BSs and femto-BSs in a 2000 \times 2000 m^2 area as shown in Fig. 3 and Fig. 4. There are 4 macro-BSs and 32 femto-BSs, where two femto-BSs are uniformly distributed in each sub-square of size 500 \times 500 m^2. There are 1000 users that arrive to the system online (one user arrival per unit time), whose locations are randomly drawn according to a non-homogeneous point process (users concentrate in interlacing sub-squares as in Fig. 3 to account for the non-uniform distribution of users in practice) and a homogeneous point process. The transmit power of a macro-BS and a femto-BS are respectively assumed to be 46 dBm and 20 dBm and the spectrum bands of the two tiers are orthogonal, each with bandwidth 10 MHz, while transmissions at the same tier interfere with each other, as has been assumed in prior work [13] and in line with industry practice. Under the massive MIMO regime, a macro-BS is assumed to have 100 antennas to provide 10 degrees of freedom for spatial multiplexing, and each macro-BS uses the same set of 10 orthogonal pilots. Similarly, a femto-BS is assumed to have 40 antennas to
Fig. 8: The effect of biasing for associating high mobility users to macro-BSs. Solid (dotted) lines are the associations for high (low) mobility users. Only a subset of users are shown.

For the case with non-homogeneous user density, Fig. 5a compares the performance of the randomized cell-centric online algorithm, the cell-centric online algorithm, the user-centric online algorithm, and the max-SINR online algorithm [12] according to which when a user arrives, the user is associated with the base station that provides the user with the highest SINR value, regardless of the system load of the base station. (Note that in the figure the sum log-rate utility is normalized with respect to the optimal value of the offline relaxation. Also, note that all the four online algorithms have similar complexity of $O(Ma)$. From Fig. 5a we observe that the sum log-rate utility of the cell-centric online algorithm is very close to the offline optimal. As a result, we do not see any performance difference between the $\frac{1}{2}$ randomized approximation algorithm and the $\frac{1}{2}$ approximation algorithm.

Motivated by the industry’s desire to offer some notion of fairness to its users, we are also interested in comparing the minimum user rates and the Jain’s fairness index [40] under the four algorithms. Note that the Jain’s fairness index is between $\frac{1}{M}$ (worst case) and 1 (best case when all users receive the same rate). As shown in Fig. 5b and Fig. 5c, the (randomized) cell-centric algorithm performs better than the others in terms of fairness too. Last, as can be seen in Fig. 5d, the max-SINR algorithm can achieve a higher sum user rate while ignoring fairness considerations. Similar results can be observed in Fig. 6 for the case with homogeneous user density.

Finally, in Fig. 7, we investigate the proportion of users associating to macro-cells and femto-cells. For both non-homogeneous and homogeneous user density cases, we observe that under the cell-centric algorithm, the proportion of users associating to femto-BSs is about 61%, while under the max-SINR algorithm, the proportion of users associating to femto-BSs is about 44%. In addition, when the transmit power of femto-BSs is increased from 20 dBm to 35 dBm, we can see that there is only a small increase in the proportion of users associating to femto-BSs (about 66% under the cell-centric algorithm and 52% under the max-SINR algorithm).

B. The effect of biasing

We investigate the effect of introducing a bias function of associating users with high mobility to tiers with large
cell coverage. In Fig. 8, we have two classes of users, high mobility and low mobility users, and two tiers consisting of either macro- or femto-cells. The proportion of users with high mobility is about 50%. Only a subset of user associations (100 out of 1000) are shown.

We consider as an example a bias function $f(x, y)$ along the lines of Fig. 2 with $x \in \{\text{high mobility, low mobility}\}$ and $y \in \{\text{macro-cell, femto-cell}\}$. Specifically, let $f(\text{high, macro}) = \gamma$, and $f(\text{high, femto}) = f(\text{low, macro}) = f(\text{low, femto}) = 0$. As we increase $\gamma$ from 0 (no bias) to 4 (a large bias), we can see from Fig. 8a–c that users with high mobility tend to be associated more and more with macro-cells.

C. Multi-channel WiFi network in MU-MIMO full multiplexing gain scenario

We consider a different network topology motivated by enterprise WiFi networks. Specifically, consider the multi-channel conference hall topology depicted in Fig. 9 and Fig. 10. There are 20 APs in a $300 \times 250$ m² area and each of them operates at one of four orthogonal channels (we use different colors to represent different channels). There are 200 users arriving to the system online (one user arrival per unit time), whose locations are independently drawn from a non-homogeneous point process (Fig. 9), i.e., a two-dimensional uncorrelated normal distribution with mean (150 m, 125 m).
and standard deviation 25 m, and a homogeneous point process
(Fig. 10). The transmit power of an AP is assumed to be 20
dBm and the channel bandwidth is assumed to be 20 MHz,
lin line with industry practice [21]. Under the MU-MIMO full
multiplexing gain regime, each AP is assumed to be equipped
with 2 antennas to provide 2 degrees of freedom for spatial
multiplexing. The noise power is $-101$ dBm, and the path loss
exponent is 3, a typical value for indoor environments [29].
For most realizations of the system deployment, and with an
SINR threshold $\tau = 3$ dB for decoding as reported in [21], the
parameter $a$ is equal to 4. For the case with non-homogeneous
user density, Fig. 11 compares the performance of the four
online algorithms in terms of the sum log-rate utility, the
minimum user rate, the Jain’s fairness index, and the sum user
rate, respectively. We can see that the (randomized) cell-centric
algorithm outperforms the others in terms of all four metrics.

Similar trends can be observed in Fig. 12 for the case with
homogeneous user density. As expected, the performance gains
are less pronounced under homogeneous user density because
users are already distributed around APs in a more balanced
way.

D. Departing users

As discussed in Section VI-B, the performance guarantee
of the randomized cell-centric algorithm holds as users arrive
online but do not leave the system. Here, we investigate the
robustness of the randomized cell-centric online algorithm
against user departures. Let us consider the topology of a two-
tier heterogeneous cellular network in Fig. 3. Suppose that
users arrive online to the system at a unit rate (one user
per time slot) from time slot 1 to time slot 1000. Upon the
arrival of a user, the user is immediately associated with one
base station (according to the randomized cell-centric online
algorithm). Starting from time slot 500, in each subsequent
time slot we randomly select one of the existing users to
leave the system (so that the number of users in the system
is maintained at 500 from time slot 500 to time slot 1000).
In Fig. 13, we compare the performance of the randomized
cell-centric online algorithm with the offline optimal, where
the offline optimal is recomputed in every time slot. We
can see that the sum utility of the randomized cell-centric
algorithm is within 1% of the offline optimal as users join
and leave the system, implying that our online algorithm is
robust against user dynamics and in practice we do not need
to re-associate users when a user departs to guarantee near
optimal performance (see discussion in Section VI-B).

VIII. Conclusion

In this paper, we proposed efficient approximation algo-
rithms for the online user association problem in a multi-
tier multi-cell mobile network, which finds applications in
todays enterprise WiFi networks and in next generation cell-
ular systems. We showed that the approximation ratio of
our champion algorithm is $\frac{4}{2a+3}$, where $a$ is the maximum
number of potential associations for a user. The parameter
$a$ is small due to the signal characteristics of the wireless
medium, and the bound constitutes a significant improvement
over the best known prior work. The proposed algorithms were
applied to many scenarios of interest including systems with
massive antenna arrays, systems with MU-MIMO capabilities,
networks with prioritized user classes, and networks where
transmitters coordinate to form clusters in a CoMP-like setup.
Last, we showed via simulations that the proposed algorithms
perform near optimal and pose desirable fairness properties
under realistic scenarios.

REFERENCES

J. Zhang, “What will 5G be?” IEEE J. Sel. Areas Commun., vol. 32,
J. Schulz, M. Samimi, and F. Gutierrez, “Millimeter wave mobile
communications for 5G cellular: It will work!” IEEE Access, vol. 1,
and analysis of K-tier downlink heterogeneous cellular networks,” IEEE
numbers of base station antennas,” IEEE Trans. Wireless Commun., vol. 9,
ing massive MIMO spectral efficiency with a not-so-large number of
3239, Sep. 2012.
[8] Y. Bejerano, S.-J. Han, and L. Li, “Fairness and load balancing in
wireless LANs using association control,” IEEE/ACM Trans. Netw.,
[9] B. Kauffmann, F. Baccelli, A. Chaintreau, V. Mhatre, K. Papagiannaki,
and C. Diot, “Measurement-based self organization of interfering 802.11
[10] Y. Bejerano and S.-J. Han, “Cell breathing techniques for load balancing
\alpha-optimal user association and cell load balancing in wireless networks,”
of load balancing in HetNets: old myths and open problems,” IEEE
APPENDIX

Suppose that $n \in \mathbb{N}, n \geq 2$ and $b_i \in \mathbb{R}, b_i > 0, i = 1, \ldots, n$. By the AM-GM inequality with $n$ variables, we have $b_1^n + (n-1)b_i^{n-1} \geq nb_1 b_i^{n-1}, i = 2, \ldots, n$. Therefore, we have

$$n \sum_{i=2}^{n} b_i b_i^{n-1} \leq \sum_{i=2}^{n} [b_1^n + (n-1)b_i^{n-1}] = (n-1) \left( b_1^n + \sum_{i=2}^{n} b_i^n \right).$$

By rearranging terms, we have

$$b_1 \left( \sum_{i=1}^{n} b_i^{n-1} \right) \leq 1 - \frac{1}{n}.$$

In general, we can conclude that

$$\frac{b_k \left( \sum_{i=1, i \neq k}^{n} b_i^{n-1} \right)}{\sum_{i=1}^{n} b_i^n} \leq 1 - \frac{1}{n}, \quad k = 1, \ldots, n.$$