An Efficient Approximation Algorithm for Online Multi-Tier Multi-Cell User Association*

Weng Chon Ao
University of Southern California
wao@usc.edu

Konstantinos Psounis
University of Southern California
kpsounis@usc.edu

ABSTRACT

The ever growing wireless bandwidth demand is pushing WiFi and cellular networks to dense multi-cell deployments, as well as to multi-tier architectures consisting of macrocells and small cells. 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 an online algorithm for the multi-tier multi-cell user association problem that has a provable performance guarantee which improves previously known bounds by a sizable amount. The proposed algorithm is motivated by online combinatorial auctions, while capturing and leveraging the relative sparsity of choices in wireless networks as compared to auction setups. Specifically, it is a $\frac{1}{2-a^{-1}}$ approximation algorithm, where a is the maximum number of feasible associations for a user and is, in general, small due to path loss. 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.

CCS Concepts

•**Networks** → *Network performance analysis; Mobile networks; Wireless local area networks;*

Keywords

User association; Load balancing; Heterogeneous networks; Online algorithm; Randomized approximation algorithm

*This work has been supported by NSF grant ECCS-1444060 and a Cisco Research Center grant.

MobiHoc'16, July 04-08, 2016, Paderborn, Germany

DOI: http://dx.doi.org/10.1145/2942358.2942360

1. INTRODUCTION

To support the tremendous growth of wireless data traffic (e.g. video streaming), a dense deployment of access points (APs) is already used in enterprise WiFi networks, 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 [18]), and the performance of the overall cellular network can be sizably improved by this heterogeneous multi-tier architecture [3, 9]. In the context of such a multi-tier, multi-cell 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, [2, 4–8, 16, 19, 23] 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 [2], while offering no performance guarantees. For example, by default, in todays cellular and WiFi networks users simply associate with the base station 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 [10], 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 a novel online algorithm for the multi-tier user-BS association problem (the single tier user-BS association problem is obviously a special case), which is both practical and provably near-optimal. The algorithm is motivated by online combinatorial auctions (bidders bid on objects) [11, 15, 17], 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 online algorithm achieves at least $\frac{1}{2-a^{-1}}$ of the offline optimal, which, for typical values of a, say 2 or 3, yields about 60 - 67% of the optimal performance. To the best of our knowledge, this is the tightest known bound achievable by online association algorithms, see Section 2 for more details.

The remainder of this paper is organized as follows. We present related work in Section 2. Section 3 describes the system model. In Section 4, we present the online multi-tier multi-cell user association algorithm. The performance analysis is given in Section 5. We discuss how the proposed algorithm can be applied in various practical scenarios of interest in Section 6. Section 7 presents numerical and simulation results for a number of real-world scenarios. Last, section 8 concludes the paper.

2. PRIOR WORK AND CONTRIBUTION

2.1 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 [2,4–8,16,19,23] 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 [7] the authors study the user-AP association problem ensuring a max-min fair bandwidth allocation. In [16] the authors perform joint AP channel selection and user association to minimize the user transmission delay. In [6], 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.

A recent overview of load balancing techniques in cellular networks can be found in [2]. In one of the works referred therein [23], 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 [8], the user-BS association problem is investigated in the context of massive MIMO enabled base stations. 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. Last, in [5, 19] the multi-tier user-BS association problem is analyzed using stochastic geometry, and in [4] a game-theoretic model is proposed to associate users with different radio access technologies.

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 reassociations to accommodate new user arrivals while maintaining high performance.

2.2 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 [20] the authors introduce a 1/8 approximation algorithm for online user-BS association to maximize the sum rate of the users under the equal time sharing scheduling, and, in [21] they introduce a 1/2 approximation algorithm to maximize the sum rate under the water-filling power allocation. Last, in [24] 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 that captures the concept of proportional fairness [23]. In addition to the fact that proportional fairness is a good approximation of the operational point of todays 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 $\frac{1}{2-a^{-1}}$ 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.

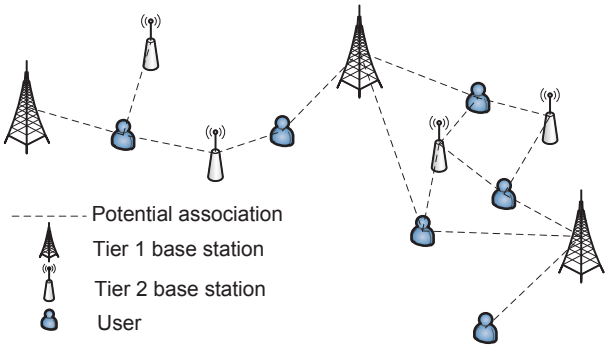


Figure 1: A scenario of multi-tier user-BS association.

3. SYSTEM MODEL

3.1 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 (i.e., the performance guarantee holds for any value of the parameter 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 $\mathcal{K} = \{1, 2, \cdots, K\}$. The bandwidth of the spectrum band of the kth tier is denoted as W_k and the spectrum bands of different tiers do not overlap. We assume that there are N_k base stations (denoted as $\mathcal{B}_k = \{1, 2, \cdots, N_k\}$) operating at tier $k \in \mathcal{K}$. As a result, each base station is indexed by a tuple (j, k), $k \in \mathcal{K}$, $j \in \mathcal{B}_k$. We consider a single carrier system where each base station in the kth 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 recourse blocks as well as a multi-channel system with pre-allocated channels, see Section 7 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.

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 τ (which is chosen to ensure successful decoding of data messages), i.e.,

$$\mathcal{A}_i \triangleq \left\{ (j,k) : \frac{P_{j,k} g_{i,j,k}}{W_k N_0 + \sum_{l \in \mathcal{B}_k, l \neq j} P_{l,k} g_{i,l,k}} \geq \tau, \right. \\ \left. k \in \mathcal{K},\ j \in \mathcal{B}_k \right\}, \quad (1)$$

where $P_{j,k}$ is the transmission power of base station j at tier k, $g_{i,j,k}$ is the channel gain between user i and base station j at tier k that captures the effects of path loss and shadowing (note that we do not consider the effect of small-scale fading since the time scale for association is much larger than that for small-scale fading), N_0 is the noise power spectral density, and $\sum_{l \in \mathcal{B}_k, l \neq j} P_{l,k} g_{i,l,k}$ is the interference received from other base stations operating at the same tier. The data rate (bits/s) between user $i \in \mathcal{U}$ and base station $(j,k) \in \mathcal{A}_i$ is given by

$$c_{i,j,k} = W_k \log \left(1 + \frac{P_{j,k} g_{i,j,k}}{W_k N_0 + \sum_{l \in \mathcal{B}_k, l \neq j} P_{l,k} g_{i,l,k}} \right), \\ i \in \mathcal{U},\ (j,k) \in \mathcal{A}_i, \quad (2)$$

where Shannon's formula is used and can be extended to accommodate real world features like modulation and coding tables, see, for example, [12].

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, equal time-sharing is used to schedule the users (the reason behind such choice is elaborated in Remark 1 at the end of Section 5). As a result, we have $r_{i,j,k} = c_{i,j,k} / \sum_{l \in \mathcal{U}} x_{l,j,k}$. Furthermore, the utility function of user i is denoted as $U_i(r_{i,j,k})$, which is a function of the actual data rate $r_{i,j,k}$. The multi-tier user-BS association problem is to find the association such that the sum utility of the users is maximized.

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 in Eq. (1) and (2) would be to replace the interference from a single tier with the sum of interference from all tiers using the same spectrum band and the analysis would work the same way.

3.2 Offline user-BS association

We first consider the offline multi-tier user-BS association problem (denoted as $\mathbf{Q}$):

$$\mathbf{Q}: \ \underset{x_{i,j,k}}{\text{maximize}} \sum_{i \in \mathcal{U}} \sum_{(j,k) \in \mathcal{A}_i} x_{i,j,k} U_i \left(\frac{c_{i,j,k}}{\sum_{l \in \mathcal{U}} x_{l,j,k}} \right) \\ \text{subject to} \sum_{(j,k) \in \mathcal{A}_i} x_{i,j,k} = 1,\ i \in \mathcal{U} \\ x_{i,j,k} \in \{0,1\}, i \in \mathcal{U},\ (j,k) \in \mathcal{A}_i, \quad (3)$$

where the constraints ensure that a user can only be associated with a single base station. We denote the value of the offline optimal as $OPT(Q)$. Since the above problem $\mathbf{Q}$ 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 the online algorithm, we derive an upper bound of $OPT(Q)$ by considering the relaxation of problem $\mathbf{Q}$ (denoted as $\tilde{\mathbf{Q}}$):

$$\tilde{\mathbf{Q}}: \ \underset{x_{i,j,k}}{\text{maximize}} \sum_{i \in \mathcal{U}} \sum_{(j,k) \in \mathcal{A}_i} x_{i,j,k} U_i \left(\frac{c_{i,j,k}}{\sum_{l \in \mathcal{U}} x_{l,j,k}} \right) \\ \text{subject to} \sum_{(j,k) \in \mathcal{A}_i} x_{i,j,k} = 1,\ i \in \mathcal{U} \\ x_{i,j,k} \geq 0, i \in \mathcal{U},\ (j,k) \in \mathcal{A}_i. \quad (4)$$

It is observed that when the utility function is $U_i(\cdot) = \log(\cdot)$ (which will be the case of our interest), the problem $\tilde{\mathbf{Q}}$ becomes a convex optimization problem and can be solved efficiently. We denote the optimal value of the problem $\tilde{\mathbf{Q}}$ as $OPT(\tilde{Q})$, and clearly we have $OPT(Q) \leq OPT(\tilde{Q})$. 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.

4. 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.

4.1 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 $\sum_{k\in\mathcal{K}}\sum_{j\in\mathcal{B}_k} s_{j,k} = M$. In practice, when a user arrives, the user can obtain the information of the system load $s_{j,k}$ by base station broadcast and the data rate $c_{i,j,k}$ by sensing and estimating the SINR, see, for example, [10]. 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 $s_{j,k} \leftarrow 0,\ k \in \mathcal{K},\ j \in \mathcal{B}_k$;
2: **for** $i = 1, \ldots, M$ **do**
3: Associate user i with base station j^* at tier k^*, where $(j^*, k^*) = \text{argmax}_{(j,k)\in\mathcal{A}_i} U_i\left(\frac{c_{i,j,k}}{s_{j,k}+1}\right)$;
4: $s_{j^*,k^*} \leftarrow s_{j^*,k^*} + 1$;
5: **end for**

4.2 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 $\mathcal{A}_{j,k}$) is the set of users that the base station j at tier k can be associated with, i.e.,

$$\mathcal{A}_{j,k} \triangleq \{i \in \mathcal{U} : (j,k) \in \mathcal{A}_i\}. \tag{5}$$

We have

$$V_{j,k}(\mathcal{S}) = \sum_{i\in\mathcal{S}} U_i\left(\frac{c_{i,j,k}}{|\mathcal{S}|}\right),\ k \in \mathcal{K},\ j \in \mathcal{B}_k,\ \mathcal{S} \subset \mathcal{A}_{j,k}, \tag{6}$$

where $\mathcal{S}$ denotes the set of users that base station (j,k) is associated with, and $|\mathcal{S}|$ is the cardinality of $\mathcal{S}$. 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 $\mathcal{S}$ as

$$V_{j,k}(i|\mathcal{S}) = V_{j,k}(\mathcal{S}\cup\{i\}) - V_{j,k}(\mathcal{S}), \quad i \in \mathcal{A}_{j,k},\ \mathcal{S} \subset \mathcal{A}_{j,k},\ i \notin \mathcal{S}. \tag{7}$$

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 $\mathcal{S}_{j,k}$ updates the set of users that base station (j,k) is associated with. At the end of the algorithm, the sets $\mathcal{S}_{j,k},\ k \in \mathcal{K},\ j \in \mathcal{B}_k$ form a partition of the users and $\sum_{k\in\mathcal{K}}\sum_{j\in\mathcal{B}_k}|\mathcal{S}_{j,k}| = M$. Also, the resulting sum utility of the users under the cell-centric randomized online algorithm (denoted as $ALG_2(Q)$) can be written as

$$ALG_2(Q) = \sum_{k\in\mathcal{K}}\sum_{j\in\mathcal{B}_k} V_{j,k}(\mathcal{S}_{j,k}), \tag{9}$$

which is the sum of a number of base station utility functions.

Algorithm 2 Cell-centric randomized online algorithm

1: Initialize $\mathcal{S}_{j,k} \leftarrow \emptyset,\ k \in \mathcal{K},\ j \in \mathcal{B}_k$;
2: **for** $i = 1, \ldots, M$ **do**
3: Associate user i with base station j at tier k with probability

$$\frac{V_{j,k}\left(i|\mathcal{S}_{j,k}\right)^{|\mathcal{A}_i|-1}}{\sum_{(j,k)\in\mathcal{A}_i} V_{j,k}\left(i|\mathcal{S}_{j,k}\right)^{|\mathcal{A}_i|-1}},\ (j,k) \in \mathcal{A}_i. \tag{8}$$

Let the selected base station be (j^*, k^*);
4: $\mathcal{S}_{j^*,k^*} \leftarrow \mathcal{S}_{j^*,k^*} \cup \{i\}$;
5: **end for**

4.3 Cell-centric deterministic online algorithm

In the previous subsection, we introduced the cell-centric randomized online algorithm. It is natural to consider its deterministic counterpart. Specifically, when a user arrives, the user is associated with the base station with the highest marginal utility (of including that user). Compared to the 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 $ALG_3(Q)$.

Algorithm 3 Cell-centric deterministic online algorithm

1: Initialize $\mathcal{S}_{j,k} \leftarrow \emptyset,\ k \in \mathcal{K},\ j \in \mathcal{B}_k$;
2: **for** $i = 1, \ldots, M$ **do**
3: Associate user i with base station j^* at tier k^*, where $(j^*, k^*) = \text{argmax}_{(j,k)\in\mathcal{A}_i} V_{j,k}\left(i|\mathcal{S}_{j,k}\right)$;
4: $\mathcal{S}_{j^*,k^*} \leftarrow \mathcal{S}_{j^*,k^*} \cup \{i\}$;
5: **end for**

5. PERFORMANCE ANALYSIS

In this section we establish the tight performance bound for the two cell-centric online algorithms. In order to apply the theory of online combinatorial auctions, we first need to prove that the specific base station utility function for our application, namely $V_{j,k}(\cdot)$ in Eq. (6), is submodular and monotone. As a concrete example, we analyze the logarithmic user utility, $U_i(\cdot) = \log(\cdot),\ \forall i \in \mathcal{U}$, which is commonly used in wireless networks to provide proportional fairness

among users [23]. Under the logarithmic user utility, the base station utility function becomes

$$V_{j,k}(\mathcal{S}) = \sum_{i\in\mathcal{S}} \log\left(\frac{c_{i,j,k}}{|\mathcal{S}|}\right),\ k\in\mathcal{K}, j\in\mathcal{B}_k, \mathcal{S}\subset\mathcal{A}_{j,k}. \quad (10)$$

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

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

LEMMA 1. *$V_{j,k}(\mathcal{S}) = \sum_{i\in\mathcal{S}}\log\left(\frac{c_{i,j,k}}{|\mathcal{S}|}\right),\ k\in\mathcal{K},\ j\in\mathcal{B}_k,\ \mathcal{S}\subset\mathcal{A}_{j,k}$ is submodular.*

PROOF. Let $i\in\mathcal{A}_{j,k},\ \mathcal{S}\subset\mathcal{T}\subset\mathcal{A}_{j,k},\ i\notin\mathcal{T}$ be given. Let us first consider the case $\mathcal{S}\neq\emptyset$. We have

$$\begin{aligned} V_{j,k}(i|\mathcal{S}) &= V_{j,k}(\mathcal{S}\cup\{i\}) - V_{j,k}(\mathcal{S}) \\ &= \sum_{l\in\mathcal{S}\cup\{i\}}\log\left(\frac{c_{l,j,k}}{|\mathcal{S}|+1}\right) - \sum_{l\in\mathcal{S}}\log\left(\frac{c_{l,j,k}}{|\mathcal{S}|}\right) \\ &= \log(c_{i,j,k}) + |\mathcal{S}|\log|\mathcal{S}| - (|\mathcal{S}|+1)\log(|\mathcal{S}|+1). \quad (11)\end{aligned}$$

Therefore, to check that $V_{j,k}(i|\mathcal{S})\geq V_{j,k}(i|\mathcal{T})$, it is equivalent to show that $|\mathcal{S}|\log|\mathcal{S}| - (|\mathcal{S}|+1)\log(|\mathcal{S}|+1) \geq |\mathcal{T}|\log|\mathcal{T}| - (|\mathcal{T}|+1)\log(|\mathcal{T}|+1)$, which in turn is equivalent to show that the function $f(x)\triangleq x\log x - (x+1)\log(x+1),\ x>0$, is decreasing. Indeed, we have

$$\begin{aligned} f'(x) &= \log x - \log(x+1) \\ &= -\log\left(1+\frac{1}{x}\right) < 0,\ \forall x>0, \quad (12)\end{aligned}$$

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

For the case with $\mathcal{S}=\emptyset$, we have $V_{j,k}(i|\mathcal{S}) = \log(c_{i,j,k})$. To check that $\log(c_{i,j,k})\geq V_{j,k}(i|\mathcal{T})$, it is equivalent to show that $0\geq|\mathcal{T}|\log|\mathcal{T}| - (|\mathcal{T}|+1)\log(|\mathcal{T}|+1)$, which in turn is equivalent to show that the function $f(x)\triangleq x\log x - (x+1)\log(x+1),\ x>0$, is less than zero. Indeed, we have

$$\begin{aligned} f(x) &= x\log x - (x+1)\log(x+1) \\ &= -\log\left[(x+1)\left(\frac{x+1}{x}\right)^x\right] \\ &= -\log\left[(1+x)\left(1+\frac{1}{x}\right)^x\right] < 0,\ \forall x>0. \quad (13)\end{aligned}$$

As a result, in all cases we have $V_{j,k}(i|\mathcal{S})\geq V_{j,k}(i|\mathcal{T})$. We conclude that $V_{j,k}(\cdot)$ is submodular. □

LEMMA 2. *If $c_{i,j,k}\geq|\mathcal{A}_{j,k}|e$ bits/s, $\forall i\in\mathcal{A}_{j,k}$, then $V_{j,k}(\mathcal{S}) = \sum_{i\in\mathcal{S}}\log\left(\frac{c_{i,j,k}}{|\mathcal{S}|}\right),\ k\in\mathcal{K},\ j\in\mathcal{B}_k,\ \mathcal{S}\subset\mathcal{A}_{j,k}$ is monotone.*

PROOF. Let $i\in\mathcal{A}_{j,k},\ \mathcal{S}\subset\mathcal{A}_{j,k},\ i\notin\mathcal{S}$ be given. From Eq. (11), we have

$$\begin{aligned} &V_{j,k}(i|\mathcal{S}) = \log(c_{i,j,k}) + |\mathcal{S}|\log|\mathcal{S}| - (|\mathcal{S}|+1)\log(|\mathcal{S}|+1) \\ &\overset{(a)}{\geq} \log c_{i,j,k} + (|\mathcal{A}_{j,k}|-1)\log(|\mathcal{A}_{j,k}|-1) - |\mathcal{A}_{j,k}|\log|\mathcal{A}_{j,k}| \\ &= \log(c_{i,j,k}) - \log\frac{|\mathcal{A}_{j,k}|^{|\mathcal{A}_{j,k}|}}{(|\mathcal{A}_{j,k}|-1)^{|\mathcal{A}_{j,k}|-1}} \\ &= \log(c_{i,j,k}) - \log|\mathcal{A}_{j,k}|\left(1+\frac{1}{|\mathcal{A}_{j,k}|-1}\right)^{|\mathcal{A}_{j,k}|-1} \\ &\geq \log(c_{i,j,k}) - \log|\mathcal{A}_{j,k}|e, \quad (14)\end{aligned}$$

where (a) holds since the function $f(x)\triangleq x\log x - (x+1)\log(x+1),\ x>0$ is decreasing and achieves its minimum when $|S| = |\mathcal{A}_{j,k}|-1$. Therefore, if $c_{i,j,k}\geq|\mathcal{A}_{j,k}|e$ bits/s, $\forall i\in\mathcal{A}_{j,k}$, we have $V_{j,k}(i|\mathcal{S})\geq 0$ and thus $V_{j,k}(\cdot)$ is monotone. □

Now, we derive the performance bound of the cell-centric randomized online algorithm.

THEOREM 1. *Under the submodularity and monotonicity of $V_{j,k}(\cdot)$, we have $E[ALG_2(Q)]\geq\frac{1}{2-a^{-1}}OPT(Q)$, where $a\triangleq\max_{i\in\mathcal{U}}|\mathcal{A}_i|$.*

Note: Recall that for the monotonicity to hold, we need $c_{i,j,k}\geq|\mathcal{A}_{j,k}|e$ bits/s, $\forall i\in\mathcal{A}_{j,k}$, i.e., we need the data rate (measured in bits/s) between base station (j,k) and user $i\in\mathcal{A}_{j,k}$ to be larger than 2.72×the number of users that base station (j,k) can be associated with, which is satisfied for any real world scenario.

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 [11], to get a lower bound equal to $\frac{1}{2-N^{-1}}\ OPT(Q)$, 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 $E[ALG_2(Q)]\geq\frac{1}{2-a^{-1}}OPT(Q)$, where $a=\max_{i\in\mathcal{U}}|\mathcal{A}_i|$ 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 $(j,k)\in\mathcal{A}_1$, we define $Q_{j,k}$ as the subproblem of associating the remaining users $2,...,M$ to the base stations, where the base station utility function $V_{j,k}(\cdot)$ is replaced by $V_{j,k}(\cdot|\{1\})$ (which is also a monotone submodular function). From the cell-centric randomized online algorithm, we have

$$E[ALG_2(Q)] = \sum_{(j,k)\in\mathcal{A}_1} q_{j,k}\left\{E[ALG_2(Q_{j,k})] + V_{j,k}(\{1\})\right\}, \quad (15)$$

where

$$q_{j,k} = \frac{V_{j,k}(\{1\})^{|\mathcal{A}_1|-1}}{\sum_{(j,k)\in\mathcal{A}_1}V_{j,k}(\{1\})^{|\mathcal{A}_1|-1}},\ (j,k)\in\mathcal{A}_1. \quad (16)$$

Let $\mathcal{S}=\{\mathcal{S}_{j,k},\ k\in\mathcal{K},\ j\in\mathcal{B}_k\}$ be the optimal offline association profile for the original problem Q and let us assume that user $1\in\mathcal{S}_{\tilde{j},\tilde{k}}$ for some $(\tilde{j},\tilde{k})\in\mathcal{A}_1$. Consider a new association profile $\mathcal{S}'$ which is the same as $\mathcal{S}$ 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 $\mathcal{S}'$ as $Val(Q_{j,k})$. (Obviously, we have $Val(Q_{j,k})\leq OPT(Q_{j,k})$.) By the submodularity and monotonicity of $V_{j,k}(\cdot)$, for all $(j,k)\in\mathcal{A}_1, (j,k)\neq(\tilde{j},\tilde{k})$, we have $OPT(Q) - Val(Q_{j,k}) \leq V_{j,k}(\{1\}) + V_{\tilde{j},\tilde{k}}(\{1\})$, where $V_{\tilde{j},\tilde{k}}(\{1\})$ is the maximum "loss" due to the fact that the subproblem $Q_{j,k}$ does not have user 1 associated with base station $(\tilde{j},\tilde{k})$, 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 $(j,k)=(\tilde{j},\tilde{k})$, we have $OPT(Q)-Val(Q_{\tilde{j},\tilde{k}})=V_{\tilde{j},\tilde{k}}(\{1\})$. As a result, we have

$$\begin{aligned}
&\frac{OPT(Q)-\sum_{(j,k)\in\mathcal{A}_1} q_{j,k}OPT(Q_{j,k})}{\sum_{(j,k)\in\mathcal{A}_1} q_{j,k}V_{j,k}(\{1\})}\\
&\leq \frac{OPT(Q)-\sum_{(j,k)\in\mathcal{A}_1} q_{j,k}Val(Q_{j,k})}{\sum_{(j,k)\in\mathcal{A}_1} q_{j,k}V_{j,k}(\{1\})}\\
&\leq \frac{\sum_{\substack{(j,k)\in\mathcal{A}_1\\(j,k)\neq(\tilde{j},\tilde{k})}} q_{j,k}[V_{j,k}(\{1\})+V_{\tilde{j},\tilde{k}}(\{1\})]+q_{\tilde{j},\tilde{k}}V_{\tilde{j},\tilde{k}}(\{1\})}{\sum_{(j,k)\in\mathcal{A}_1} q_{j,k}V_{j,k}(\{1\})}\\
&= 1+\frac{V_{\tilde{j},\tilde{k}}(\{1\})\sum_{(j,k)\in\mathcal{A}_1,(j,k)\neq(\tilde{j},\tilde{k})} V_{j,k}(\{1\})^{|\mathcal{A}_1|-1}}{\sum_{(j,k)\in\mathcal{A}_1} V_{j,k}(\{1\})^{|\mathcal{A}_1|}}\\
&\overset{(a)}{\leq} 1+1-\frac{1}{|\mathcal{A}_1|}\leq 2-\frac{1}{\max_{i\in\mathcal{U}}|\mathcal{A}_i|}=2-\frac{1}{a}, \qquad (17)
\end{aligned}$$

where (a) follows by the AM-GM inequality (see Appendix).

Therefore, we have

$$\begin{aligned}
OPT(Q) &\overset{(a)}{\leq} \sum_{(j,k)\in\mathcal{A}_1} q_{j,k}OPT(Q_{j,k})\\
&\qquad +\left(2-\frac{1}{a}\right)\sum_{(j,k)\in\mathcal{A}_1} q_{j,k}V_{j,k}(\{1\})\\
&\overset{(b)}{\leq} \sum_{(j,k)\in\mathcal{A}_1} q_{j,k}\left(2-\frac{1}{a}\right)[E[ALG_2(Q_{j,k})]+V_{j,k}(\{1\})]\\
&\overset{(c)}{=} \left(2-\frac{1}{a}\right)E[ALG_2(Q)], \qquad (18)
\end{aligned}$$

where (a) follows from Eq. (17), (b) follows by induction, and (c) follows from Eq. (15). □

Note that the above performance bound holds for a generic submodular and monotone base station utility function $V_{j,k}(\cdot)$, while the logarithmic user utility function has been used in Eq. (10) to serve as an example.

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 $ALG_3(Q)\geq\frac{1}{2}OPT(Q)$.*

Note: Recall that for the monotonicity to hold, we need $c_{i,j,k}\geq|\mathcal{A}_{j,k}|e$ bits/s, $\forall i\in\mathcal{A}_{j,k}$.

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 [17]. □

REMARK 1. *Optimality of equal time allocation*: In the above analysis, we assume that equal time sharing is used to schedule transmissions for users associated with the same base station (see Eq. (10)). 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 equal time sharing is optimal.

For any base station $(j,k),\ k\in\mathcal{K},\ j\in\mathcal{B}_k$, let $\mathcal{S}\subset\mathcal{A}_{j,k}$ be the set of users associated with it. Similarly to [23], let us define the time sharing variables $w_{i,j,k},\ i\in\mathcal{S}$ where $\sum_{i\in\mathcal{S}} w_{i,j,k}=1$ and $w_{i,j,k}\geq 0,\ i\in\mathcal{S}$. The time sharing variables are optimized such that the sum utility of the users in $\mathcal{S}$ is maximized. In other words, the base station utility function is generalized from Eq. (10) to

$$\begin{aligned}
V_{j,k}(\mathcal{S}) = \underset{w_{i,j,k}}{\text{maximize}}\ &\sum_{i\in\mathcal{S}}\log\left(w_{i,j,k}c_{i,j,k}\right)\\
\text{subject to}\ &\sum_{i\in\mathcal{S}} w_{i,j,k}=1\\
&w_{i,j,k}\geq 0,\ i\in\mathcal{S}. \qquad (19)
\end{aligned}$$

It is not hard to see that the optimal time sharing variables are $w^*_{i,j,k}=\frac{1}{|\mathcal{S}|}, i\in\mathcal{S}$, showing that equal time allocation is optimal for the logarithmic user utility.

6. PRACTICAL CONSIDERATIONS

In the following, we consider scenarios of practical interest and show how the proposed cell-centric randomized online algorithm can be applied into these scenarios.

6.1 Base stations with multiple antennas

When a base station has multiple antennas (without loss of generality, the users are still assumed to be equipped with a single antenna), precoding such as zero-forcing beamforming (ZFBF) [22] can be used at the base station to support multiple simultaneous data transmissions/streams to its associated users. These data transmissions are spatially isolated and will not interfere with each other. For any base station $(j,k),\ k\in\mathcal{K},\ j\in\mathcal{B}_k$, suppose that the base station has $\eta_{j,k}$ antennas (to provide $\eta_{j,k}$ degrees of freedom [22]) and allocates power equally on each data stream. The data rate (bits/s) between base station (j,k) and user $i\in\mathcal{A}_{j,k}$ is thus given by

$$c_{i,j,k}=W_k\log\left(1+\frac{\frac{P_{j,k}}{\eta_{j,k}}g_{i,j,k}}{W_kN_0+\sum_{l\in\mathcal{B}_k,l\neq j}P_{l,k}g_{i,l,k}}\right). \qquad (20)$$

Let $\mathcal{S}\subset\mathcal{A}_{j,k}$ be the set of users associated with base station (j,k). The multi-antenna base station utility function is written as

$$V_{j,k}(\mathcal{S})=\begin{cases}\sum_{i\in\mathcal{S}}\log\left(c_{i,j,k}\right) & \text{if } |\mathcal{S}|\leq\eta_{j,k}\\ \sum_{i\in\mathcal{S}}\log\left(\frac{\eta_{j,k}c_{i,j,k}}{|\mathcal{S}|}\right) & \text{if } |\mathcal{S}|>\eta_{j,k}.\end{cases} \qquad (21)$$

When the number of the associated users $|\mathcal{S}|$ is less than or equal to the number of base station antennas (the degrees of freedom) $\eta_{j,k}$, each user can be active for the whole duration without time sharing by using ZFBF. However, time sharing is still needed when $|\mathcal{S}|>\eta_{j,k}$. Last, the marginal multi-antenna base station utility function can be derived as

$$\begin{aligned}
&V_{j,k}(i|\mathcal{S})=\log(c_{i,j,k}), \quad \text{if } |\mathcal{S}|\leq\eta_{j,k}-1;\\
&V_{j,k}(i|\mathcal{S})=\log(c_{i,j,k})+\log(\eta_{j,k})+|\mathcal{S}|\log|\mathcal{S}|\\
&\qquad -(|\mathcal{S}|+1)\log(|\mathcal{S}|+1), \quad \text{if } |\mathcal{S}|\geq\eta_{j,k}. \qquad (22)
\end{aligned}$$

LEMMA 3. *The multi-antenna base station utility function $V_{j,k}(\cdot)$ in Eq. (21) is submodular. In addition, if $c_{i,j,k}\geq\max\left\{\frac{|\mathcal{A}_{j,k}|e}{\eta_{j,k}},1\right\}$ bits/s, $\forall i\in\mathcal{A}_{j,k}$, $V_{j,k}(\cdot)$ is monotone.*

PROOF. Let $i\in\mathcal{A}_{j,k},\ \mathcal{S}\subset\mathcal{T}\subset\mathcal{A}_{j,k},\ i\notin\mathcal{T}$ be given. There are three cases. In the first case with $|\mathcal{T}|\leq\eta_{j,k}-1$, we clearly have $V_{j,k}(i|\mathcal{S})=V_{j,k}(i|\mathcal{T})$. In the second case

with $|\mathcal{S}| \leq \eta_{j,k} - 1$ and $|\mathcal{T}| \geq \eta_{j,k}$, we have

$$\begin{aligned} & V_{j,k}\left(i|\mathcal{S}\right) - V_{j,k}\left(i|\mathcal{T}\right) \\ &= -\log(\eta_{j,k}) - |\mathcal{T}|\log|\mathcal{T}| + (|\mathcal{T}|+1)\log(|\mathcal{T}|+1) \\ &= \log\frac{(|\mathcal{T}|+1)^{|\mathcal{T}|+1}}{|\mathcal{T}|^{|\mathcal{T}|}\eta_{j,k}} \geq \log\frac{(|\mathcal{T}|+1)^{|\mathcal{T}|+1}}{|\mathcal{T}|^{|\mathcal{T}|+1}} > 0. \end{aligned} \quad (23)$$

In the third case with $|\mathcal{S}| \geq \eta_{j,k}$, we have $V_{j,k}\left(i|\mathcal{S}\right) \geq V_{j,k}\left(i|\mathcal{T}\right)$ by a similar argument as in the proof of Lemma 1. Last, we can check the condition for the monotonicity of $V_{j,k}(\cdot)$ as in the proof of Lemma 2. □

With the submodularity and monotonicity of $V_{j,k}(\cdot)$, the performance guarantee of Algorithm 2 and 3 can be applied to the multi-antenna case as well.

6.2 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. 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.

6.3 User dynamics

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 $\mathcal{S}^{m-1} \triangleq \left\{\mathcal{S}_{j,k}^{m-1},\ k \in \mathcal{K},\ j \in \mathcal{B}_k\right\}$ generated at the $m-1^{\text{th}}$ iteration of Algorithm 2) and then re-associate users $m+1$ to M. Clearly, this may result in a number of re-asscoaitions, which is not practical. In Section 7, 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.

7. SIMULATION RESULTS

7.1 Two-tier heterogeneous cellular network

We consider a two-tier heterogeneous cellular network consisting of macro-BSs and femto-BSs in a 2000×2000 m^2 area as shown in Fig. 2 and Fig. 3. There are 4 macro-BSs and 32 femto-BSs, where two femto-BSs are uniformly distributed in each sub-square of size 500×500 m^2. There are 840 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. 2 to account for the non-uniform

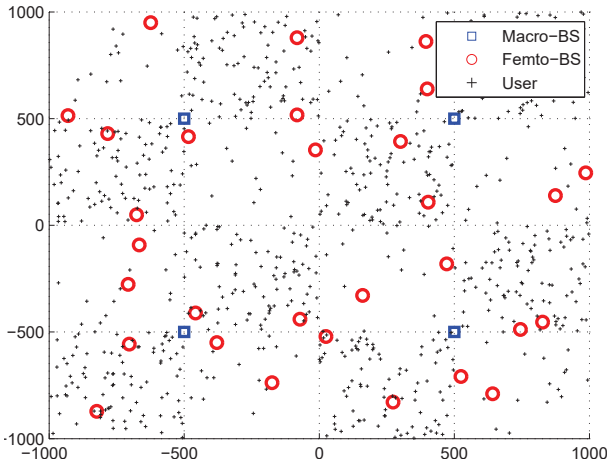


Figure 2: Two-tier heterogeneous network with non-homogeneous user density.

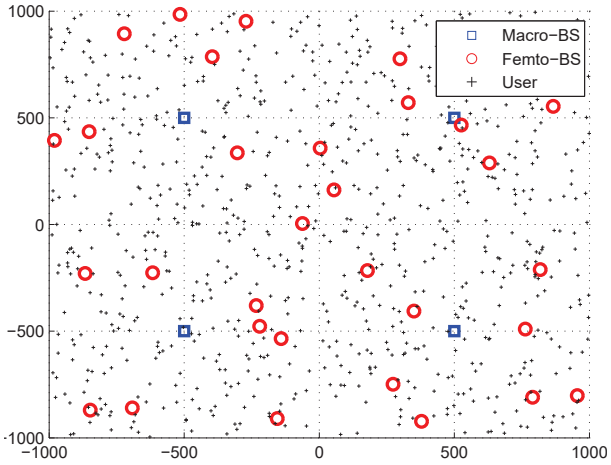
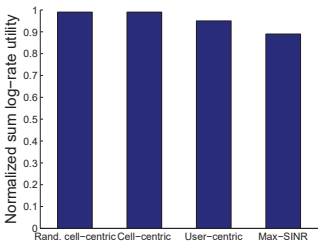


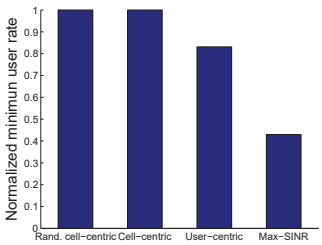
Figure 3: Two-tier heterogeneous network with homogeneous user density.

distribution of users in practice) and a homogeneous point process (Fig. 3). 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 [23] and in line with industry practice. The background noise power is assumed to be -104 dBm, and the path loss exponent is supposed to be 4, as is usually the case in outdoor environments [12]. Last, given the above parameters, for most realizations of the system deployment, and with an SINR threshold $\tau = -3$ dB for decoding as measured in real-world deployments [13], the maximum number of potential associations of a user is equal to $a = 3$.

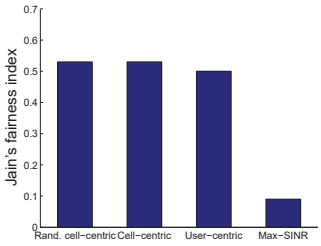
For the case with non-homogeneous user density, Fig. 4a compares the performance of the randomized cell-centric online algorithm, the cell-centric online algorithm, the user-centric online algorithm, and the max-SINR online algorithm [2] according to which when a user arrives, the user is associated 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



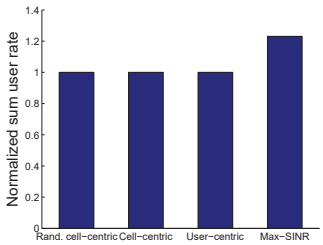
(a) The sum log-rate utility (normalized w.r.t. the offline optimal).



(b) The minimum user rate (normalized w.r.t. the rand. cell-centric).

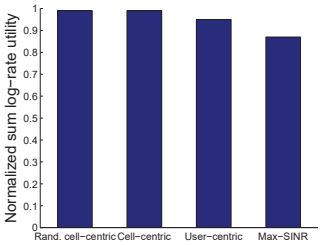


(c) The Jain's fairness index.

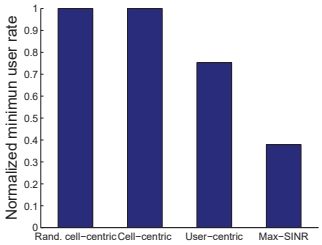


(d) The sum user rate (normalized w.r.t. the rand. cell-centric).

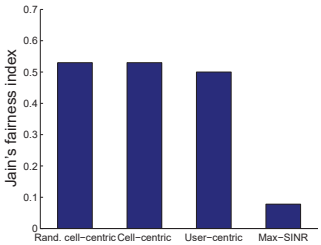
Figure 4: Performance under the two-tier heterogeneous network with non-homogeneous user density.



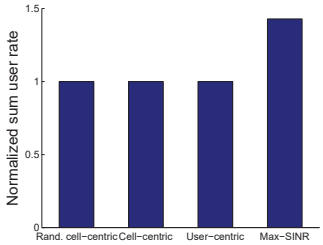
(a) The sum log-rate utility (normalized w.r.t. the offline optimal).



(b) The minimum user rate (normalized w.r.t. the rand. cell-centric).



(c) The Jain's fairness index.



(d) The sum user rate (normalized w.r.t. the rand. cell-centric).

Figure 5: Performance under the two-tier heterogeneous network with homogeneous user density.

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. 4a, 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 $\frac{1}{2-a^{-1}}$ 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 [14] 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. 4b and Fig. 4c, the (randomized) cell-centric algorithm performs better than the others in terms of fairness too. Last, as can be seen in Fig. 4d, the max-SINR algorithm can achieve a higher sum user rate while ignoring the user fairness.

Similar results can be observed in Fig. 5 for the case with homogeneous user density. Also, the results remain similar as we vary other system level parameters, e.g. transmit power and noise levels, spectrum bandwidth, path loss exponent and other channel characteristics, within reasonable ranges found in real world scenarios.

7.2 Multi-channel WiFi network

We consider a different network topology motivated by enterprise WiFi networks. Specifically, consider the multi-channel conference hall topology depicted in Fig. 6 and Fig. 7. There are 20 APs in a 300×250 m^2 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

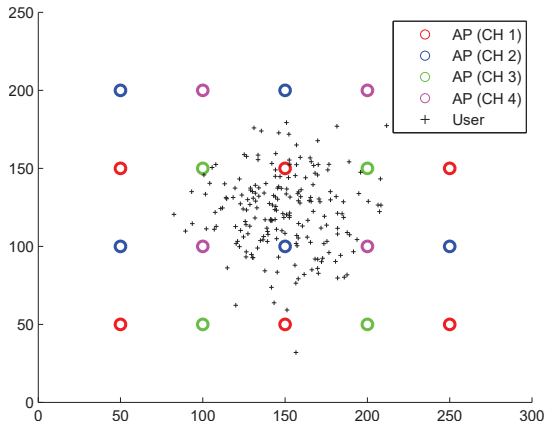
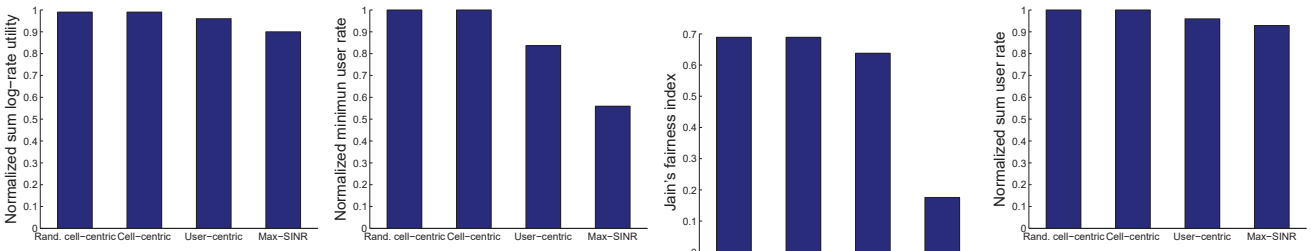


Figure 6: Multi-channel WiFi network with non-homogeneous user density.

time), whose locations are independently drawn from a non-homogeneous point process (Fig. 6), 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. 7). The transmit power of an AP is assumed to be 20 dBm and the channel bandwidth is assumed to be 20 MHz, in line with industry practice [10]. The noise power is -101 dBm, and the path loss exponent is 3, a typical value for indoor environments [12]. For most realizations of the system deployment, and with an SINR threshold $\tau = 3$ dB for decoding as reported in [10], the parameter a is equal to 4. For the case with non-homogeneous user density, Fig. 8 compares the performance of the four online algorithms in



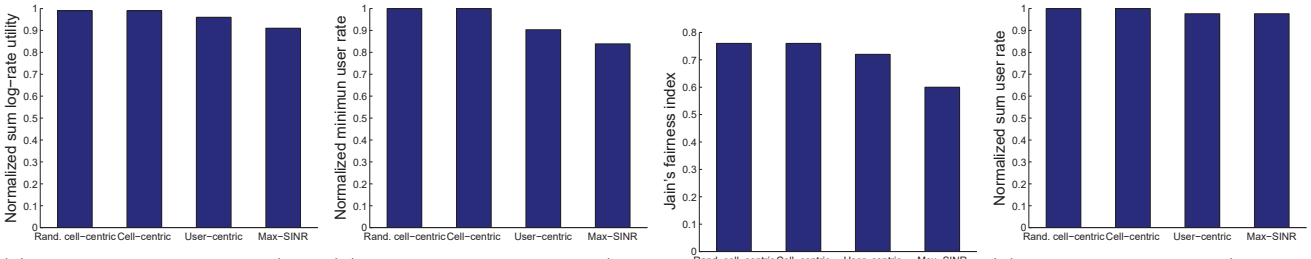
(a) The sum log-rate utility (normalized w.r.t. the offline optimal).

(b) The minimum user rate (normalized w.r.t. the rand. cell-centric).

(c) The Jain's fairness index.

(d) The sum user rate (normalized w.r.t. the rand. cell-centric).

Figure 8: Performance under the multi-channel WiFi network with non-homogeneous user density.



(a) The sum log-rate utility (normalized w.r.t. the offline optimal).

(b) The minimum user rate (normalized w.r.t. the rand. cell-centric).

(c) The Jain's fairness index.

(d) The sum user rate (normalized w.r.t. the rand. cell-centric).

Figure 9: Performance under the multi-channel WiFi network with homogeneous user density.

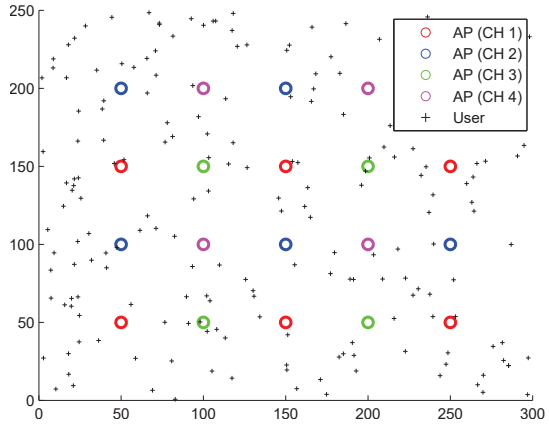


Figure 7: Multi-channel WiFi network with homogeneous user density.

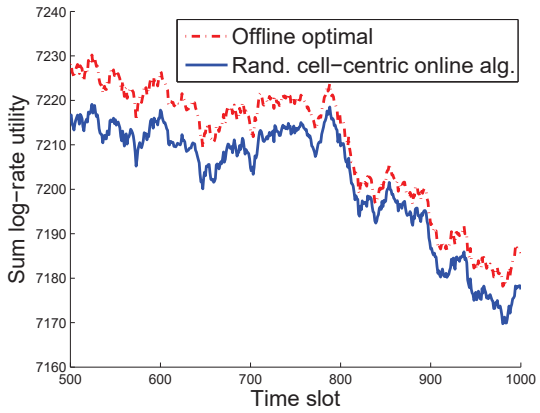


Figure 10: Performance of the randomized cell-centric online algorithm against user dynamics.

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. 9 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.

7.3 User dynamics

As discussed in Section 6.3, 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 dynamics. Let us consider the topology of a two-tier heterogeneous cellular network in Fig. 2. 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. 10, 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 (see discussion in Section 6.3).

8. CONCLUSION

In this paper, we proposed an efficient approximation algorithm for the online multi-tier multi-cell user association problem, which finds applications in todays enterprise WiFi networks and in next generation cellular systems. We showed that the approximation ratio of the proposed algorithm is $\frac{1}{2-a^{-1}}$, 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. In addition, we showed via simulations that the proposed algorithm performs near optimal and poses desirable fairness properties under realistic scenarios.

9. REFERENCES

[1] J. Andrews, S. Buzzi, W. Choi, S. Hanly, A. Lozano, A. Soong, and J. Zhang. What will 5G be? *IEEE J. Sel. Areas Commun.*, 32(6):1065–1082, June 2014.

[2] J. Andrews, S. Singh, Q. Ye, X. Lin, and H. Dhillon. An overview of load balancing in HetNets: old myths and open problems. *IEEE Wireless Commun. Mag.*, 21(2):18–25, Apr. 2014.

[3] W. C. Ao and K. Psounis. Distributed caching and small cell cooperation for fast content delivery. In *Proc. ACM MobiHoc*, 2015.

[4] E. Aryafar, A. Keshavarz-Haddad, M. Wang, and M. Chiang. RAT selection games in HetNets. In *Proc. IEEE INFOCOM*, Apr. 2013.

[5] W. Bao and B. Liang. Structured spectrum allocation and user association in heterogeneous cellular networks. In *Proc. IEEE INFOCOM*, Apr. 2014.

[6] Y. Bejerano and S.-J. Han. Cell breathing techniques for load balancing in wireless LANs. *IEEE Trans. Mobile Comput.*, 8(6):735–749, June 2009.

[7] Y. Bejerano, S.-J. Han, and L. Li. Fairness and load balancing in wireless LANs using association control. *IEEE/ACM Trans. Netw.*, 15(3):560–573, June 2007.

[8] D. Bethanabhotla, O. Y. Bursalioglu, H. C. Papadopoulos, and G. Caire. Optimal user-cell association for massive MIMO wireless networks. *IEEE Trans. Wireless Commun.*, 15(3):1835–1850, Mar. 2016.

[9] S. M. Cheng, W. C. Ao, F. M. Tseng, and K. C. Chen. Design and analysis of downlink spectrum sharing in two-tier cognitive femto networks. *IEEE Trans. Veh. Technol.*, 61(5):2194–2207, June 2012.

[10] Cisco. Cisco Wireless LAN Controller Configuration Guide. Technical report, Cisco Systems, Inc., 2013.

[11] S. Dobzinski and M. Schapira. An improved approximation algorithm for combinatorial auctions with submodular bidders. In *Proc. ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2006.

[12] A. Goldsmith. *Wireless Communications*. Cambridge University Press, 2005.

[13] H. Holma, A. Toskala, and J. Reunanen. *LTE Small Cell Optimization: 3GPP Evolution to Release 13*. Wiley, 2016.

[14] R. Jain, D. Chiu, and W. Hawe. A quantitative measure of fairness and discrimination for resource allocation in shared computer systems. *DEC Research Report TR-301*, Sept. 1984.

[15] M. Kapralov, I. Post, and J. Vondrák. Online submodular welfare maximization: Greedy is optimal. In *Proc. ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2013.

[16] B. Kauffmann, F. Baccelli, A. Chaintreau, V. Mhatre, K. Papagiannaki, and C. Diot. Measurement-based self organization of interfering 802.11 wireless access networks. In *Proc. IEEE INFOCOM*, May 2007.

[17] B. Lehmann, D. Lehmann, and N. Nisan. Combinatorial auctions with decreasing marginal utilities. *Games and Economic Behavior*, 55(2):270–296, 2006.

[18] T. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. Wong, J. Schulz, M. Samimi, and F. Gutierrez. Millimeter wave mobile communications for 5G cellular: It will work! *IEEE Access*, 1:335–349, May 2013.

[19] S. Singh, H. Dhillon, and J. Andrews. Offloading in heterogeneous networks: Modeling, analysis, and design insights. *IEEE Trans. Wireless Commun.*, 12(5):2484–2497, May 2013.

[20] A. Thangaraj and R. Vaze. Online algorithms for basestation allocation. *IEEE Trans. Wireless Commun.*, 13(5):2966–2975, May 2014.

[21] K. Thekumparampil, A. Thangaraj, and R. Vaze. Combinatorial resource allocation using submodularity of waterfilling. *IEEE Trans. Wireless Commun.*, 15(1):206–216, Jan. 2016.

[22] D. Tse and P. Viswanath. *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

[23] 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.*, 12(6):2706–2716, June 2013.

[24] Y. Zhang, D. Bethanabhotla, T. Hao, and K. Psounis. Near-optimal user-cell association schemes for real-world networks. In *Proc. ITA*, 2015.

APPENDIX

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

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

By rearranging terms, we have

$$\frac{b_1(\sum_{i=2}^{n} b_i^{n-1})}{\sum_{i=1}^{n} b_i^n} \leq 1 - \frac{1}{n}.$$

In general, we can conclude that

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