IEEE TRANSACTIONS ON
EMERGING TOPICS
IN COMPUTING

Received 30 June 2014; revised 17 November 2014; accepted 6 January 2015.
Date of publication 27 January, 2015; date of current version 4 September, 2015.

Digital Object Identifier 10.1109/TETC.2015.2395093

Approximation Algorithms for Cell Association and Scheduling in Femtocell Networks

HUI ZHOU, SHIWEN MAO (Senior Member, IEEE), AND PRATHIMA AGRAWAL (Fellow, IEEE)

Department of Electrical and Computer Engineering, Auburn University, Auburn, AL 36849 USA

This work was supported in part by the U.S. National Science Foundation (NSF) under Grant CNS-1247955.

ABSTRACT Femtocells are recognized as effective for improving network coverage and capacity and for reducing power consumption due to the reduced range of wireless transmissions. Although highly appealing, a plethora of challenging problems need to be addressed for fully harvesting its potential. In this paper, we investigate the problem of cell association and service scheduling in femtocell networks. In addition to the general goal of offloading traffic from the macrobase station (BS), we also aim at minimizing the latency of service requested by the users, while considering both open and closed access strategies. We show that the cell association problem is NP-hard, and propose several near-optimal solution algorithms for assigning users to BSs, including a sequential fixing algorithm, a rounding approximation algorithm, a greedy approximation algorithm, and a randomized algorithm. For service scheduling, we develop an optimal algorithm to minimize the average waiting time for the users associated with the same BS. The proposed algorithms are analyzed with respect to performance bounds, approximation ratios, and optimality, and are evaluated with simulations.

INDEX TERMS Approximation algorithm, cell association, femtocell, load balancing, randomized algorithm.

I. INTRODUCTION

The current and next generation mobile communication networks are facing the grand challenge of 1000 times wireless data increase by 2020 as compared to the 2010 level [1]. As indicated in the Qualcomm report, more spectrum and small cells are the key elements in the solution to meet this challenge. With the massive bandwidth in the mmWave band, many bandwidth-demanding new applications can be easily supported in mmWave wireless networks [2]–[4]. On the other hand, small cells are the enabler of efficient spatial reuse of the spectrum resource, which has achieved the largest increase in wireless network capacity in the past decades, comparing to other technical advances such as advanced modulation and coding techniques [5].

In this paper, we consider the problem of cell association and service scheduling in femtocell networks. A femtocell, as shown in Fig. 1, is a relatively small cellular network with a femtocell base station (FBS), usually deployed in places where signal reception from the macro base station (MBS) is weak due to long distance or obstacles. An FBS is typically the size of a residential gateway or even smaller and connects to the service provider's network via broadband connections. FBS is designed to serve approved users within its coverage to offload wireless traffic from the MBS. Due to shortened wireless transmission range, femtocell is shown effective in reducing transmit power and boosting signal-to-interference-plus-noise ratio (SINR), which lead to prolonged battery life of mobile devices, improved network coverage, and enhanced network capacity [5].

Femtocells have gained a lot of attention from both academia and industry in the recent past. The three largest cellular network operators in the United States (i.e., AT&T, Sprint and Verizon) have all offered commercial femtocell products and service recently. Although highly promising, a plethora of problems with both technical and economic natures have not been fully addressed yet. In [5], a discussion is provided on the challenging technical issues in femtocell networks, ranging from synchronization, cell association, network organization, to quality of service (QoS) provisioning.

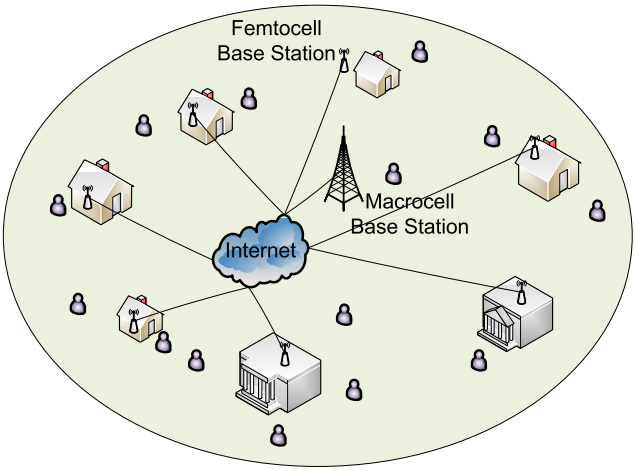


FIGURE 1. Illustration of a two-tier femtocell network.

Unlike the MBS, whose placement is planned and optimized by operators, FBS's are usually randomly deployed by users. When the chaotic femtocell placement meets randomly distributed mobile users, cell association (or load balancing) becomes a critical problem for the performance of femtocell networks. For example, an FBS might be deployed at a place with high user density. With an inappropriate cell association strategy, this FBS may have to serve all the users within its coverage, leading to very high load at this FBS and high service latency for its users. An effective cell association scheme should be used in this case to evenly distribute the load among neighboring FBS's and/or the MBS. The cell association problem is particularly prominent in femtocell networks due to the unreliability of FBS's. The operation of an FBS may be interrupted by its owner (e.g., turned off after office hours); it may also experience power outage or any other faults. Then all the users initially associated with this FBS should be quickly assigned to other neighboring FBS's or the MBS. It is a load balancing problem on how to effectively associate these users with neighboring BS's without introducing a load burst and performance degradation at a particular BS.

In this paper, we investigate the problem of cell association and service scheduling in a two-tier femtocell network, as shown in Fig. 1. In addition to the general goal of offloading wireless traffic from the MBS, we also aim to minimize the latency of service requested by users, while considering both open and closed access strategies. In particular, we consider one MBS and multiple FBS's serving randomly distributed mobile users. Users request to the BS's for downlink transmission of data packets. Without loss of generality, we assume that each user is allowed to connect to either the MBS or an FBS. The cell associate problem is to assign the users to the BS's such that the transmission of all the data packets can be completed as soon as possible. When multiple users are associated with one BS, we also aim to develop a service scheduling scheme such that the average waiting time for the users will be minimized.

We provide a general framework for the cell association problem for both open and closed access scenarios, which can be reduced to the classic load balancing problem and is NP-hard [6]. Therefore, we aim to develop effective near-optimal algorithms with guaranteed performance. In particular, we first provide a *sequential fixing algorithm* based on a linear programming (LP) relaxation, which can achieve the best performance among the proposed schemes but with a relatively high computational complexity [7]. To reduce the complexity, we propose a *rounding approximation algorithm* that ensures a $(\rho + 1)$-approximation of the optimal solution, and a *greedy approximation algorithm* that ensures a (2κ)-approximation of the optimal solution, where ρ and κ are threshold parameters that will be introduced in Sections IV-C1 and IV-C2, respectively. To further reduce the requirement on frequently updated channel state information (CSI), we then develop a *randomized algorithm* that allows a user to randomly pick a BS from a reduced BS list to connect to. Once the reduced BS list is generated by the randomized algorithm, no information exchange is required among users. An *upper bound* for the maximum expected service time achieved by the randomized algorithm is then derived. After the users are assigned to the BS's, we next address the service scheduling problem for determining the transmission order of the data packets requested by the users associated with the same BS. We develop a simple algorithm to minimize the average waiting time for the users, and prove its *optimality*.

In addition to rigorous analysis of the proposed algorithms with respect to performance bounds, approximation ratios, and optimality, we also evaluate the proposed schemes with simulations, where superior performance is observed. It is worth noting that although the algorithms are developed in the context of femtocell networks, they should also be applicable to the case of small cells such as picocells, where the more coordination among the BS's can be exploited to make the algorithms more effective.

The remainder of this paper is organized as follows. The related work is discussed in Section II. We present the system model in Section III, and the problem formulation and proposed algorithms in Section IV. The scheduling problem is presented in V. The proposed algorithm are evaluated in Section VI. Section VII concludes this paper.

II. RELATED WORK

Femtocells have been acknowledged as an effective solution to the capacity crisis of wireless networks. Ref. [5] provided comprehensive discussions of the technical issues, regulatory concerns, and economic incentives in femtocell networks. Generally speaking, there are three different access control strategies in femtocell networks, open access, closed access and hybrid access. The pros and cons of these strategies were studied in [8].

Deploying femtocells also means introducing interference if no appropriate mitigation strategy is incorporated. Considerable research have been conducted on interference

mitigation by assigning users to proper orthogonal channels [9].

Apart from the studies on interference mitigation, there are an increasing number of papers on cell association or cell selection under various scenarios [10]–[16]. Dhahri and Ohtsuki in [10] proposed a learning-based cell selection method for an open access femtocell network. Madan et al. in [11] described new paradigms of cell association in heterogeneous networks (HetNet) with the help of third-party backhaul connections. Their simple and lightweight methodologies and algorithms incur a low signaling overhead. In [12], a convex optimization problem was formulated for cell association and a dynamic range extension algorithm was proposed to maximize the minimum rate of users on the downlink of the HetNet. However, this paper did not directly optimize the load balancing in the HetNet, but rather focused on the sum rate and minimum rate. In [13], a cell association and access control scheme was presented to maximize network capacity while achieving fairness among users. In [14], the authors provided an analytic framework for evaluating outage probability and spectral efficiency with flexible cell association in heterogeneous cellular networks. Mukherjee in [15] analyzed the downlink SINR distribution in HetNet with biased cell association. We also developed centralized and distributed algorithms for cell association in a massive MIMO enabled HetNet in a recent work [16].

There are also some interesting prior work on load balancing in cellular networks. A theoretical framework was presented in [17] for distributed user association and cell load balancing under spatially heterogeneous traffic distribution. A distributed α-optimal algorithm was proposed and it supports different load-balancing objectives, which include rate-optimal, throughput-optimal, delay-optimal, and load-equalizing, as α is set to different values. In [18], the authors developed an off-line optimal algorithm for load balancing to achieve network-wide proportional fairness in multi-cell networks. They considered partial frequency reuse (PFR) jointly with load-balancing in a multi-cell network to achieve network-wide proportional fairness. A practical on-line algorithm was also proposed and the expected throughput was taken as the decision making metric. On-line assignment when users arrive one at a time was studied extensively in computer science literature. The competitive ratio analysis in [19] showed that any deterministic on-line algorithm can achieve a competitive ratio of $\log n$, where n is the number of servers.

We find most of the related research was focused on offloading MBS traffic and improving network capacity with FBS's. In the following sections, we propose several cell association and transmission scheduling schemes with the objective of minimizing service latency in femtocell networks.

III. SYSTEM MODEL

We consider a two-tier femtocell network with M base stations: one MBS (indexed by 1) and $M-1$ FBS's (indexed from 2 to M). All the BS's are connected to the Internet via broadband wired connections. There are N mobile users randomly located within the coverage of the femtocell network. We assume the MBS and FBS's are well synchronized and share the same spectrum. We consider Time Division Multiplexing (TDM) systems. Without loss of generality, we assume each user requests a fixed-length data packet from one of the M BS's. The problem is to assign the users to the BS's and schedule the transmission of their requested data packets at each BS, such that the transmissions can be finished as earlier as possible.

A. DOWN LINK CAPACITY

Let P_m be the transmit power of BS m and $G_{m,n}$ the channel gain between the BS and user n. According to the Shannon Theorem, the link capacity of user n connected to BS m is given by

$$C_{m,n} = B\log_2\left(1+\frac{G_{m,n}P_m}{I_{m,n}+\sigma^2}\right), \tag{1}$$

where B is network bandwidth, $I_{m,n}$ is the interference from all other BS's, and σ^2 is the noise power density. It follows that

$$I_{m,n} = \sum_{i=1}^{M} G_{i,n}P_i - G_{m,n}P_m = I_n - G_{m,n}P_m, \tag{2}$$

where I_n is the sum of interference from all BS's to user n. It does not depend on which BS user n is connected to and is a constant for each user. Substituting (2) into (1), we have

$$\begin{aligned} C_{m,n} &= B\log_2\left(1+\frac{G_{m,n}P_m}{I_n - G_{m,n}P_m+\sigma^2}\right) \\ &= B\log_2\left(\frac{1}{1-\eta_{m,n}}\right), \end{aligned} \tag{3}$$

where $\eta_{m,n}$ is the signal to interference plus noise ratio (SINR), the same ratio of the received power in I_n at user n.

As can be seen in (3), CSI is required to compute the downlink capacities. We assume the BS's keep on measuring the CSIs of the downlink channels. Channel reciprocity could be exploited to simplify channel estimation. When a user m moves, the impact is a change in the channel gain between this user and each BS.

B. SERVICE TIME

Without loss of generality, we assume each user requests a fixed-length data packet from one of the BS's. For simplicity of notation, we assume all the packets have the same length, denoted as L. Then the transmission (or, service) time at BS m for user n is given by

$$t_{m,n} = L/C_{m,n}. \tag{4}$$

The service time depends on the link capacity $C_{m,n}$ as given in (3) and the packet length L. Note that the service time defined here is actually the transmission delay, i.e., the

time it takes to finish the transmission of the data packet. The propagation delay is negligible due to the short distance and is ignored.

C. FEMTOCELL ACCESS CONTROL

The type of access control for femtocells can be classified into two categories: closed access and open access. The *open-access strategy* allows all mobile users of an operator to connect to any of the FBS's. In this case, femtocells are often deployed by an operator to enhance coverage in an area where there is a coverage hole. With the *closed-access strategy*, only a specific user group can get service from the FBS's [20]. Although closed access has been shown to reduce the system throughput by 15%, surveys suggest that closed access is users' favorite option [21].

In this paper, we consider both access strategies. Let $\mathcal{A}_m$ denote the set of users that can connect to BS m and $\mathcal{B}_n$ the set of BS's that user n can connect to. Both open and closed access strategies can be easily modeled by these two sets. Specifically, for open access, we have $\mathcal{A}_m = \{1, \ldots, N\}$ and $\mathcal{B}_n = \{1, \ldots, M\}$.

IV. CELL ASSOCIATION PROBLEM FORMULATION AND PROPOSED ALGORITHMS

To make the complex problem tractable, we divide the problem into two steps. First, we assign each user to one of the M BS's with the objective of minimizing the total service time on each BS. Second, we schedule the service order for the set of users associated with the same BS to minimize the average waiting time of users.

A. PROBLEM STATEMENT

The cell association problem can be formulated as a load balancing problem. Given a set of N users and a set of M BS's. Each user n has a service time $t_{m,n}$ if it is connected to BS m. Let $\mathcal{C}_m$ denote the set of users assigned to BS m. Then it takes a total amount of time $T_m = \sum_{n\in\mathcal{C}_m} t_{m,n}$ for BS m to transmit all the packets. For optimal network-wide performance, we seek to minimize the maximum load among all the BS's, i.e.,

$$\min\ T = \max_m\{T_m\} = \max_m \left\{ \sum_{n\in\mathcal{C}_m} t_{m,n} \right\}. \tag{5}$$

We find the cell association problem is similar to a load balancing problem. However, our problem is more challenging than the classic load balancing problem, where the service time of a user is identical when connecting to any BS. In our cell associate problem, the service time is a function of the link capacity as in (4). Its solution depends on not only user n, but also BS m. This cell association problem is easily seen to be NP-hard: when all the $t_{m,n}$'s are identical for any BS m, the problem is reduced to the classic load balancing problem, which is NP-hard [6].

In the remainder of this section, we develop effective algorithms to solve the cell association problem. In particular, we present a sequential fixing algorithm, two approximation algorithms, as well as a randomized algorithm, and derive the associated approximation ratios and performance bounds. The complexity and performance of the proposed algorithms are summarized in Table 1.

TABLE 1. Comparison of the proposed algorithms.

Algorithm	*Complexity*	*Performance*
Sequential Fixing	$O((MN)^{4.5}L_b)$	n/a
Rounding Approx.	$O((MN)^{3.5}L_b)$	$(\rho+1)$-approximation
Greedy Approx.	$O(MN)$	(2κ)-approximation
Randomized	$O(MN^2)$	Upper bound (19)
Service Scheduling	$\mathcal{O}(K\log K)$	optimal

B. SEQUENTIAL FIXING ALGORITHM

To solve the above problem, we first define indicator variables $x_{m,n}$ as

$$x_{m,n} = \begin{cases} 1, & \text{if user } n \text{ is connected to BS } m \\ 0, & \text{otherwise.} \end{cases} \quad \text{for all } m, n. \tag{6}$$

Then we reformulate the problem as follows.

$$\begin{aligned} &\min\ T && && && (7)\\ &\text{s.t.} \sum_m x_{m,n} = 1, && \text{for all } n \\ &\quad \sum_n t_{m,n}x_{m,n} \le T, && \text{for all } m \\ &\quad x_{m,n} \in \{0,1\}, && \text{for all } n\in\mathcal{A}_m, && \text{for all } m \\ &\quad x_{m,n} = 0, && \text{for all } n\notin\mathcal{A}_m, && \text{for all } m. \end{aligned}$$

In the formulated problem (7), all the indicator variable $x_{m,n}$'s are binary, while T is a real variable. Thus it is a mixed integer linear programming problem [6], denoted by MILP, which is usually NP-hard.

The original MILP is next relaxed to a linear programming (LP) problem, denoted as RLP. Specifically, we allow binary variable $x_{m,n}$'s to take real values in [0, 1]. The MILP problem is relaxed into RLP as follows:

$$\begin{aligned} &\min\ \ T && && && (8)\\ &\text{s.t.}\ \sum_m x_{m,n} = 1, && \text{for all } n \\ &\quad \sum_n t_{m,n}x_{m,n} \le T, && \text{for all } m \\ &\quad x_{m,n} \ge 0, && \text{for all } n\in\mathcal{A}_m, && \text{for all } m \\ &\quad x_{m,n} = 0, && \text{for all } n\notin\mathcal{A}_m, && \text{for all } m. \end{aligned}$$

Since the sum of $x_{m,n}$'s is already upper bounded by 1 in the first constraint, we remove the upper bound 1 of $x_{m,n}$'s in the third constraint. Obviously, the solution to the RLP problem is a lower bound of the original MILP problem because it is obtained by expanding the solution space. Unfortunately, it is usually an infeasible solution to the original MILP problem. Therefore, we develop a sequential fixing (SF) algorithm [7], [22] to find a feasible solution to the MILP problem, which is presented in Algorithm 1.

Algorithm 1: Sequential Fixing for Cell Association

```
Initialize 𝒩 = {1, ..., N};
Relax x_{m,n} to real numbers;
while 𝒩 is not empty do
    Solve the RLP problem;
    Find x_{m',n'} that is the closest to the nearest integer:
    {m', n'} = arg min_{m∈ℬ_n, n∈𝒜_m∩𝒩} {x_{m,n}, 1 − x_{m,n}} ;
    Set x_{m',n'} to the closest integer;
    if x_{m',n'} is set to 1 then
        Set x_{m,n'} = 0 for all m ≠ m';
        Remove n' from 𝒩;
    else
        Remove n' from A_{m'};
    end
end
```

In Algorithm 1, we solve the RLP problem iteratively. During each iteration, we find the $x_{m',n'}$ that has the minimum value for $(x_{m,n}-0)$ or $(1-x_{m,n})$ among all the fractional $x_{m,n}$'s, and round it up or down to the nearest integer. Setting $x_{m',n'}$ to 1 means user n' is connected to BS m'. Therefore, user n' cannot be connected to any other BS's and the rest of $x_{m,n'}$'s are set to 0, for all $m \neq m'$. This procedure repeats until all the $x_{m,n}$'s are fixed to either 0 or 1.

The complexity of SF depends on the specific LP algorithm. With Karmarkar's algorithm, the worst-case polynomial bound for solving LP problems is $O({n_v}^{3.5}L_b)$, where n_v is the number of variables and L_b is the number of bits of input to the algorithm. We have the following proposition.

Proposition 1: The computational complexity of the sequential fixing algorithm is $O((MN)^{4.5}L_b)$.

Proof: The number of binary variables in MILP is at most MN, so the number of loops in sequential fixing problem is at most MN. In each iteration, the complexities of Steps 4, 5 and the rest of the steps are $O((MN)^{3.5}L_b)$, $O(MN)$ and $O(1)$, respectively. Besides, in each iteration, the number of variables is reduced by 1. Therefore, the complexity of SF is given by

$$\sum_{i=1}^{MN} O((MN-i+1)^{3.5}L_b) = \sum_{i=1}^{MN} O(i^{3.5}L_b) = O((MN)^{4.5}L_b).$$

Therefore, the complexity of SF is upper bounded by $O((MN)^{4.5}L_b)$. □

C. APPROXIMATION ALGORITHMS

Although the sequential fixing algorithm can solve the MILP problem within polynomial time, its complexity may be high even for moderately sized femtocell networks. In this section, we propose an approximation algorithm with low complexity to solve the MILP problem. Before we introduce the approximation algorithm, we first present the following lemma on T^*.

Lemma 1: The optimal solution, denoted by T^**, to the MILP problem is lower bounded by* $T^* \geq \frac{1}{M}\sum_{n=1}^{N} \underline{t}_n$ *where* $\underline{t}_n = \min_{m\in\mathcal{B}_n} t_{m,n}$.

Proof: Given the optimal allocation $\mathcal{C}_m^*$ for BS m, for all m, we have $T^* = \max_m \sum_{n\in\mathcal{C}_m^*} t_{m,n}$. It follows that

$$T^* \geq \max_m \sum_{n\in\mathcal{C}_m^*} \underline{t}_n \geq \frac{1}{M}\sum_{m=1}^{M}\sum_{n\in\mathcal{C}_m^*} \underline{t}_n = \frac{1}{M}\sum_{n=1}^{N} \underline{t}_n.$$

The first inequality is due to the definition of $\underline{t}_n$. The second inequality is due to the fact that the maximum value is always greater than the mean value. The last equality is because all users have to be connected to one of the BS's and $\cup_{m=1}^{M}\mathcal{C}_m^*$ is the set of all users. □

Furthermore, the maximum total service time is at least the service time of any one user. We then have the following lemma on T^*.

Lemma 2: The optimal solution, denoted by T^**, to the MILP problem is lower bounded by* $T^* \geq \max \underline{t}_n$*, where* $\underline{t}_n = \min_{m\in\mathcal{B}_n} t_{m,n}$.

These lemmas will be used in analyzing the approximation ratio of the proposed approximation algorithms, which are presented in following subsections.

1) ROUNDING APPROXIMATION ALGORITHM

To ensure the required SINR for each user, $\mathcal{B}_n$ should not include all the FBS's in a real femtocell network. For example, some faraway FBS should not be considered by a user. Thus, we can use a threshold ρ to obtain the subsets $\mathcal{A}_m$ and $\mathcal{B}_n$ ($\mathcal{A}_m$ will be updated when $\mathcal{B}_n$ is determined).

$$\mathcal{B}_n' = \mathcal{B}_n \cap \left(\left\{m \,\middle|\, \frac{t_{m,n}}{\underline{t}_n} \leq \rho\right\}\right), \quad \mathcal{A}_m' = \{n | m \in \mathcal{B}_n'\}. \quad (9)$$

Usually only a limited number of FBS's will be taken into consideration for a user, due to the small coverage of femtocells. After we adopt this threshold, not only users' SINR requirements will be satisfied, but also the computational complexity will be greatly reduced.

Once $\mathcal{A}_m'$ and $\mathcal{B}_n'$ are determined, the following relaxed LP problem can be solved by an LP solver.

$$\begin{aligned} \min \quad & T \\ \text{s.t.} \quad & \sum_m x_{m,n} = 1, \quad \text{for all } n \\ & \sum_n t_{m,n}x_{m,n} \leq T, \quad \text{for all } m \\ & x_{m,n} \geq 0, \quad \text{for all } n \in \mathcal{A}_m', \quad \text{for all } m \\ & x_{m,n} = 0, \quad \text{for all } n \notin \mathcal{A}_m', \quad \text{for all } m. \end{aligned} \quad (10)$$

We denote the solution obtained by solving this RLP program by T. Since now the x-variables are allowed to take fractional values, we have $T \leq T^*$.

Without sequentially fixing these fractional values, we adopt a rounding method from [23] to obtain a feasible solution for the MILP problem. In this rounding method, a *bipartite graph* is constructed according to the RLP solution, which is constructed as a undirected bipartite

graph $G(\mathcal{A}\cup\mathcal{B}, E)$. In the disjoint set $\mathcal{A}$, each node represents a user n, while the other disjoint set $\mathcal{B}$ consists of BS nodes. We create $k_m = \lceil \sum_n x_{m,n} \rceil$ nodes in $\mathcal{B}$ for BS m and these node are denoted by $\{b_{m,1}, b_{m,2}, \ldots, b_{m,k}, \ldots, b_{m,k_m}\}$. The edges are determined in the following way. For BS m, we sort the users in the order of non-increasing service time $t_{m,n}$ and the users are renamed $\{u_1, u_2, \cdots\}$. Let $X_{m,u_j} = \sum_{i=1}^{j} x_{m,u_i}$. For each BS m, we divide the users associated to it into k_m groups, as $G_1, G_2, \cdots, G_{k_m}$. User u_j will be included in group k $(1 \le k \le k_m)$ if $k-1 < X_{m,u_j} \le k$ or $k-1 \le X_{m,u_{j-1}} < k$. If a user u_j is included in two groups, the association x-variables need to be adjusted, such that $x'_{b_{m,k},u_j} = X_{m,u_j} - k + 1$ and $x'_{b_{m,k-1},u_j} = x_{m,u_j} - x'_{b_{m,k},u_j}$. Then we insert edges between BS node $b_{m,k}$ and all the user nodes in group k. Now the bipartite graph is created and we next find a *maximum matching* $\mathcal{M}$ from each user to nodes in the other disjoint set. This maximum matching $\mathcal{M}$ indicates a feasible solution for the MILP problem: for each edge $(n, b_{m,k})$ in $\mathcal{M}$, we associate user n to BS m.

Let $T_{(b_{m,k})}$ denote the total service time at node $b_{m,k}$ before the matching operation and $T'_{(b_{m,k})}$ the total service time at node $b_{m,k}$ obtained by the above rounding method. We have the following lemma.

Lemma 3: For each node $b_{m,k}$, where $k_m \ge k > 1$, we have $T_{(b_m,k-1)} \ge T'_{(b_m,k)}$.

Proof: First, observe that the minimum service time in group $(k-1)$ will always be no less than the maximum service time in group k, because we sort the users according to their service times in the non-increasing order.

According to the above bipartite graph construction, for any $k < k_m$, we have $\sum_{i\in G_k} x'_{b_{m,k},u_i} = 1$; for $k = k_m$, we have $\sum_{i\in G_k} x'_{b_{m,k},u_i} \le 1$.

$T'_{(b_m,k)}$ will be no greater than the maximum service time in group k and will thus be no greater than the minimum service time in group $(k-1)$, which is less than $\sum_{i\in G_{k-1}} x'_{b_{m,k-1},u_i} t_{m,u_i}$. Since $T_{(b_m,k-1)} = \sum_{i\in G_{k-1}} x'_{b_{m,k-1},u_i} t_{m,u_i}$, consequently, we have the conclusion that $T_{(b_m,k-1)} \ge T'_{(b_m,k)}$. □

Now we show that the solution produced by the rounding approximation algorithm is at most $(\rho+1)$ times greater than the optimal solution.

Theorem 1: The approximation algorithm based on linear programming and the rounding method ensures a $(\rho+1)$-approximation of the optimal solution.

Proof: For each BS m, we create k_m nodes for it and there are k_m corresponding groups of user nodes adjacent to them. Thus the total service time is $\sum_{k=1}^{k_m} T'_{(b_m,k)}$.

According to Lemma 3, we have $T_{(b_m,k-1)} \ge T'_{(b_m,k)}$ for $k_m \ge k > 1$. It follows that

$$\sum_{k=2}^{k_m} T'_{(b_m,k)} \le \sum_{k=1}^{k_m-1} T_{(b_m,k)} \le \sum_{k=1}^{k_m} T_{(b_m,k)} \le T.$$

In the first group, the maximum load will be the maximum service time of users associated with m. According to Lemma 2 and the definition of ρ in (9), we have $T'_{(b_m,1)} \le \max t_{m,n} \le \rho \max \underline{t}_n \le \rho T^*$. Then, the total service time on any BS computed by our association algorithm will be $\sum_{k=1}^{k_m} T'_{(b_m,k)} \le \rho T^* + T \le (\rho+1)T^*$. The last inequality was due to $T \le T^*$, since T is the solution of the relaxed problem (10). The proof is completed. □

The complexity to compute a maximum matching is $O(VE)$, where V and E are the number of nodes and edges, respectively. Since we only need to run the matching algorithm once to obtain the association relationship, the total computational complexity of this algorithm is $O((MN)^{3.5}L_b)$, which is lower than that of SF.

Proposition 2: The computational complexity of the rounding approximation algorithm is $O((MN)^{3.5}L_b)$.

2) GREEDY APPROXIMATION ALGORITHM

We next present a low complexity approximation algorithm, where the BS with the lowest load is greedily chosen and the user whose completion time at this BS is the smallest is assigned to this BS.

We define $\kappa_{m,n} = t_{m,n}/\underline{t}_n$ and

$$\kappa = \max_{\{m,n\}} \kappa_{m,n},$$

which will be used in the optimality analysis. The greedy approximation algorithm is presented in Algorithm 2. In Step 4, we find the candidate BS for users that has the minimum T_m. Then we pick the user who has the minimum $T_{m,n}$ at the chosen BS in Step 5. Obviously, the computational complexity of the approximation algorithm is $O(MN)$, which is much lower than that of sequential fixing.

Algorithm 2: Greedy Approximation Algorithm for Cell Association

```
1  Initialize T_m = 0 and C_m = φ for all BS's ;
2  Set the user set N = {1, ..., N} ;
3  while N is not empty do
4      Find the BS m' that has the minimum T_m:
       m' = arg min_{m∈(∪_{n∈N} B_n)} {T_m} ;
5      Find the user n' that has the minimum t_{m',n}:
       n' = arg min_{n∈{A_{m'} ∩ N}} {t_{m',n}} ;
6      Set C_{m'} = C_{m'} ∪ {n'} ;
7      Set T_{m'} = T_{m'} + t_{m',n'} ;
8      Set κ_{m',n'} = t_{m',n'}/t_{n'} ;
9      Remove n' from N ;
10 end
```

Proposition 3: The computational complexity of the greedy approximation algorithm is $O(MN)$.

We have the following lemma for the performance of the greedy approximation algorithm.

Lemma 4: The greedy approximation algorithm solution, denoted by T, is upper bounded by $\frac{\kappa}{M}\sum_{n=1}^{N} \underline{t}_n + \kappa \cdot T^*$.

Proof: We first consider the open access strategy where each user can connect to any of the BS's. In the l-th iteration in Algorithm 2, we choose the BS with the minimum T_m

in Step 4. Thus we have

$$T_{m'}^{(l-1)} \le \frac{1}{M}\sum_{m=1}^{M} T_m^{(l-1)} = \frac{1}{M}\sum_{m=1}^{M}\sum_{n\in\mathcal{C}_m^{(l-1)}} t_{m,n}$$
$$= \frac{1}{M}\sum_{m=1}^{M}\sum_{n\in\mathcal{C}_m^{(l-1)}} \kappa_{m,n}\cdot \underline{t}_n \le \frac{\kappa^{(l-1)}}{M}\sum_{m=1}^{M}\sum_{n\in\mathcal{C}_m^{(l-1)}} \underline{t}_n,$$

where $\kappa^{(l-1)} = \max_{\{m,n\in\mathcal{C}_m^{(l-1)}\}} \kappa_{m,n}$. Note that $\mathcal{C}_m^{(l-1)}$ is the set of users that have been assigned to BS m in the $(l-1)$th iteration.

In Step 5, we pick user n' and let user n' connect to BS m'. Since $\kappa^{(l)}$ will always be greater than $\kappa^{(l-1)}$ and according to Lemma 2, we have

$$T_{m'}^{(l-1)} + t_{m',n'} \le \frac{\kappa^{(l)}}{M}\sum_{m=1}^{M}\sum_{n\in\mathcal{C}_m^{(l)}} \underline{t}_n + \kappa^{(l)}\underline{t}'_n.$$

The algorithm steps after N iterations. Since $T^{(l+1)} = \max\{T^{(l)}, T_{m'}^{(l)} + t_{m',n'}\}$ and $T^{(0)} = 0$, we conclude that

$$T = T^{(N+1)} = \max\left\{T^{(N)}, T_{m'}^{(N)} + t_{m',n'}\right\}$$
$$\le \frac{\kappa}{M}\sum_{m=1}^{M}\sum_{n\in\mathcal{C}_m} \underline{t}_n + \kappa\cdot T^* = \frac{\kappa}{M}\sum_{n=1}^{N}\underline{t}_n + \kappa\cdot T^*.$$

With the closed access strategy, we set $t_{m,n} = \infty$ for BS m that user n cannot connect to, for all m, n. The proof follows the same procedure and we have the same conclusion. □

Combining Lemmas 1 and 4, we have the following theorem regarding the performance of Algorithm 2.

Theorem 2: The greedy approximation algorithm in Algorithm 2 ensures a (2κ)-approximation of optimal solution.

Proof: The proof is straightforward. We have

$$T^* \le T \le \frac{\kappa}{M}\sum_{n=1}^{N}\underline{t}_n + \kappa\cdot T^* \le 2\kappa\cdot T^*,$$

where T^* is the optimal solution and T is the greedy approximation algorithm solution. Note that unlike in Section IV-C1, we have $T^* \le T$ since there is no relaxation here. □

From Theorem 2, κ is an important parameter to the performance of the greedy approximation algorithm. The smaller the κ, the smaller the optimality gap. In order to make the greedy approximation algorithm solution more competitive, we only allow users to choose from a subset $\mathcal{B}_n$ of the original BS set. Then we have the new subsets $\mathcal{B}'_n$ and $\mathcal{A}'_m$ as

$$\mathcal{B}'_n = \mathcal{B}_n \cap \left(\left\{m \,\middle|\, \frac{t_{m,n}}{\underline{t}_n} \le \Gamma\right\} \cup \{1\}\right), \mathcal{A}'_m = \{n | m \in \mathcal{B}'_n\}, \tag{11}$$

where Γ is a predefined threshold and $\{1\}$ is the index of the MBS. Γ can also be used to indicate the SINR requirement of users. The set $\mathcal{A}_m$ is replaced by $\mathcal{A}'_m$ accordingly. This way, the greedy approximation algorithm solution is bounded as given in the following Corollary.

Corollary 2.1: When Γ is used as a threshold in determining subsets $\mathcal{B}'_n$ and $\mathcal{A}'_m$ as in (11), the greedy approximation algorithm solution is bounded as

$$T^* \le T \le 2\Gamma T^*. \tag{12}$$

D. RANDOMIZED ALGORITHM

Both the rounding and greedy approximation algorithms are centralized algorithms that require frequent CSI updates. In this section, we introduce a randomized algorithm for the cell association problem. With the randomized algorithm, each user n randomly chooses a subset of $\mathcal{B}_n$ to connect to. Once the subsets are determined, no information exchange is required among the users. We assume user n connects to BS m with probability $p_{m,n}$ and the expected service time for user n on each BS is identical (i.e., by tuning the $p_{m,n}$'s), i.e.,

$$p_{m,n}\cdot t_{m,n} = H_n, \text{ for all } m \in \mathcal{B}_n.$$

Since a BS with a smaller $t_{m,n}$ should have a higher preference, we set $p_{m,n}$ proportional to $1/t_{m,n}$. Since each user has to choose a BS to connect to, we have $\sum_{m\in\mathcal{B}_n} p_{m,n} = 1$ for all n. It follows that

$$H_n = \frac{1}{\sum_{m\in\mathcal{B}_n} 1/t_{m,n}}, \text{ for all } n. \tag{13}$$

The expected load on BS m, denoted by $\bar{T}_m$, is

$$\bar{T}_m = \mathbb{E}[T_m] = \sum_{n\in\mathcal{A}_m} t_{m,n}\cdot p_{m,n} = \sum_{n\in\mathcal{A}_m} H_n, \quad \text{for all } m. \tag{14}$$

Since users are randomly connected to the BS's, our objective is to minimize the maximum value of the expected load $\bar{T}_{max}$.

$$\min\ \bar{T}_{max} = \min\left\{\max_m\left\{\bar{T}_m\right\}\right\}. \tag{15}$$

It can be seen from (14) that minimizing $\bar{T}_m$ is equivalent to reducing the number of users in $\mathcal{A}_m$.

The randomized algorithm consists of two phases. In Phase I, we use a threshold Λ to obtain the subsets $\mathcal{A}_m$ and $\mathcal{B}_n$, as

$$\mathcal{B}'_n = \mathcal{B}_n \cap (\{m | t_{m,n} \le \Lambda\} \cup \{1\}), \mathcal{A}'_m = \{n | m \in \mathcal{B}'_n\}. \tag{16}$$

Note that the subsets $\mathcal{A}'_m$ and $\mathcal{B}'_n$ are different from those defined in (11): Λ is the upper bound of service time $t_{m,n}$, while Γ is the upper bound on the service time ratios. Thus we have all $t_{m,n} \le \Lambda$ for all n and $n \in \mathcal{A}'_m$. Then we derive the upper bounds for H_n, $\bar{T}_m$ and $\bar{T}_{max}$ as

$$\begin{cases} H_n = \frac{1}{\sum_{m\in\mathcal{B}'_n} 1/t_{m,n}} \le \frac{1}{\sum_{m\in\mathcal{B}'_n} 1/\Lambda} = \frac{\Lambda}{|\mathcal{B}'_n|} \\ \bar{T}_m = \sum_{n\in\mathcal{A}'_m} H_n \le \frac{|\mathcal{A}'_m|}{\min_n |\mathcal{B}'_n|}\Lambda \\ \bar{T}_{max} = \max_m \bar{T}_m \le \frac{\max_m |\mathcal{A}'_m|}{\min_n |\mathcal{B}'_n|}\Lambda. \end{cases} \tag{17}$$

where $|\mathcal{A}'_m|$ and $|\mathcal{B}'_n|$ are the cardinalities of subsets $\mathcal{A}'_m$ and $\mathcal{B}'_n$, respectively.

In Phase II, we aim to further reduce the sizes of $\mathcal{A}'_m$ and $\mathcal{B}'_n$. From (13), we find that $H_{n'}$ gets increased when BS m' is removed from set $\mathcal{B}'_{n'}$ and user n' is removed from

set $\mathcal{A}'_{m'}$ simultaneously. The amount of increase, denoted by $\Delta_{m',n'}$, is given by

$$\Delta_{m',n'} = \frac{1}{\sum_{m\in\mathcal{B}'_n} 1/t_{m,n} - 1/t_{m',n'}} - \frac{1}{\sum_{m\in\mathcal{B}'_n} 1/t_{m,n}} = \frac{1/t_{m',n'}}{(\sum_{m\in\mathcal{B}'_n} 1/t_{m,n} - 1/t_{m',n'})(\sum_{m\in\mathcal{B}'_n} 1/t_{m,n})}. \quad (18)$$

For those BS's in the set $\{m|m \in \mathcal{B}'_{n'}, m \neq m'\}$, their $\bar{T}_m$'s become larger when BS m' is removed from set $\mathcal{B}'_{n'}$ and user n' is removed from set $\mathcal{A}'_{m'}$. On the other hand, $\bar{T}_{m'}$ is reduced by $H_{m',n'}$ according to (14).

The randomized algorithm is presented in Algorithm 3. In Step 2, we find the users that each has more than one BS on their BS list $\mathcal{B}'_n$. Then from Step 5 to Step 18, we find the BS m' with the largest $\bar{T}_{m'}$ and compute the possible maximum load $\bar{T}^{max}_{m',n}$ on BS's for all users that might be connected to BS m', assuming user n is removed from $\mathcal{A}''_{m'}$. In Step 19, we pick user n' with the minimum $\bar{T}^{max}_{m',n}$ value. If the value is less than the original $\bar{T}_{m'}$, we remove the BS-user pair $\{m', n'\}$ from sets $\mathcal{A}''_{m'}$ and $\mathcal{B}''_{n'}$. Otherwise, the algorithm is terminated. When the algorithm is executed, sets $\mathcal{A}''_{m'}$ and $\mathcal{B}''_{n'}$ are subsets of $\mathcal{A}'_{m'}$ and $\mathcal{B}'_{n'}$, respectively. Since the complexity from Step 5 to Step 18 is $O(MN)$ in the worst case, the complexity of the entire randomized algorithm is $O(MN^2)$.

Proposition 4: The computational complexity of the randomized algorithm is $O(MN^2)$.

Finally, we have the following theorem on the performance of the randomized algorithm.

Theorem 3: The maximum expected service time achieved by the randomized algorithm is upper bounded by

$$\bar{T}_{max} \leq \frac{\max_m |\mathcal{A}''_m|}{\min_n |\mathcal{B}''_n|} \times \max_n \max_{m\in\mathcal{B}''_n} t_{m,n}. \quad (19)$$

Proof: The proof is similar to the derivation of (17), but the new upper bound of service time, $\max_n \max_{m\in\mathcal{B}''_n} t_{m,n}$, is used, instead of the service time bound Λ. □

Algorithm 3: Randomized Algorithm for Cell Association

```
1  Initialize A''_m = A'_m, B''_n = B'_n ;
2  Set the user set N = {n||B''_n| > 1} ;
3  Compute T̄_m according to (14) ;
4  while N is not empty do
5      Find the BS m' with m' = arg max_m T̄_m ;
6      for user n in (A''_{m'} ∩ N) do
7          Compute Δ_{m',n} according to (18) ;
8          for m = 1 to M do
9              if m = m' then
10                 Set T̄'_{m'} = T̄_{m'} − H_n ;
11             else if m in {m|m ∈ B''_n} then
12                 Set T̄'_m = T̄_m + Δ_{m',n} ;
13             else
14                 Set T̄'_m = T̄_m ;
15             end
16         end
17         Set T̄^{max}_{m',n} = max_m T̄'_m ;
18     end
19     Find user n' with n' = arg min_n T̄^{max}_{m',n} ;
20     if T̄_{m'} ≥ T̄^{max}_{m',n'} then
21         Remove m' from B''_{n'} and n' from A''_{m'} ;
22         Update all T̄_m's ;
23         if |B''_{n'}| = 1 then
24             Remove n' from N ;
25         end
26     else
27         The algorithm is terminated ;
28     end
29 end
```

V. SERVICE SCHEDULING

Once the cell associate problem is solved as in Section IV, we then investigate how to schedule the transmissions of multiple users connecting to the same BS. Since we assume the bandwidth B is fully utilized for transmitting a user's data packet (i.e., TDD systems. See (3)), the packets are transmitted consecutively. We need to determine the service order of the users that are associated with the same BS.

Consider a tagged BS to which K users are connected. The user service times are $\{t_1, t_2, \ldots, t_K\}$. If the service order follows the user index, the average waiting time is given by

$$\bar{T}_{wait} = \frac{1}{K}\sum_{n=1}^{K}\sum_{i=1}^{n} t_i. \quad (20)$$

We have the following theorem to minimize the average waiting time $\bar{T}_{wait}$.

Theorem 4: Given K users with service times $\{t_1, t_2, \cdots, t_K\}$, the average waiting time is minimized when the users are served in the increasing order of their service times.

Proof: First, we sort the users according to their service times in the increasing order. The ordered service times are denoted by $\{t'_1, \ldots, t'_K\}$. Consider two ordered users i and j, where $1 \leq i < j \leq K$. We have $t'_i \leq t'_j$. If the positions of i and j are swapped, it is obvious that the waiting times of users from 1 to $i-1$ and the users from j to K are not affected and remain the same values, respectively. However, the awaiting time for each user from i to $j-1$ is increased by $t'_j - t'_i$. Therefore, we conclude that the average waiting time is minimized when the users are served in the increasing order of their service times. □

The complexity of the service scheduling algorithm is the same as sorting the K service times, i.e., $O(K \log K)$.

VI. PERFORMANCE EVALUATION

In this section, we evaluate the performance of the proposed cell association and service scheduling algorithms using MATLAB simulations. The channel models from [24] are adopted in our simulations. The channel gain (in dB) from the BS's to users can be expressed as $10\log(G_{m,n}) = -PL_m(d_{m,n}) - u_m$, where $d_{m,n}$ is the distance from BS m to user n, and u_m is the shadowing effect, which is normally distributed with a zero mean and variance δ_m. The simulation parameters are presented in Table 2. In the figures, each point is the average of 10 simulation runs with different

TABLE 2. Simulation parameters.

Paramter	*Value*
Number of BS's	6
Total network bandwidth	10 MHz
Transmit power of the MBS	43 dBm
Transmit power of the FBS	31.5 dBm
Path loss model for MBS	$28 + 35\log_{10}(d)$
Path loss model for FBS	$38.5 + 20\log_{10}(d)$
Shadowing effect	6 dB
Packet length	1 KBytes
Threshold ρ	5

random seeds. We plot 95% confidence intervals as error bars to make the simulation results credible.

We present simulation results for the following two scenarios: (i) open access femtocells; (ii) closed access femtocells. For comparison purpose, we also developed and simulated a *selfish scheme* and compared it with the proposed schemes. With the selfish scheme, every user simply chooses the BS with the best channel condition to connect to.

We found the sequential fixing algorithm has a high computation complexity and may not be suitable for practical systems. However, we still include this scheme in the simulation studies as a benchmark scheme for demonstrating the performance of other schemes.

A. OPEN ACCESS STRATEGY

In the first scenario, there are $M = 6$ BS's, i.e., one MBS and five FBS's. The number of users ranges from 30 to 80 with step size 10. They are randomly located in network area. Each user can connect to one of the BS's.

We first examine the impact of the number of users on total service time. In Fig. 2(b), we plot the maximum total service time for the five algorithms along with the lower bound found by solving the relaxed LP. As expected, the more users, the more total service time on BS's. Except for the low bound, the sequential fixing algorithm achieves the smallest total service time. The rounding approximation algorithm has a slightly better performance than the greedy approximation algorithm and the result justifies the approximation ratio proven in Section IV-C. Both approximation algorithms always achieve lower load than both the randomized algorithm and the selfish scheme. We also observe that beyond 50 users, all the proposed algorithms have lower service times than the simple selfish scheme. When number of users becomes larger, the simple selfish scheme becomes less competitive and the rounding approximation algorithm achieves almost 50% less total service time in the case of 80 users.

After cell association, users should be properly scheduled to get service in BS's to minimize average waiting time. In Fig. 2(b), we investigate the impact of the number of users on average waiting time. In the scheme of greedy approximation, randomized algorithm and sequential fixing, we use the service scheduling policy in Section V to schedule users in BS's and obtain the corresponding waiting time.

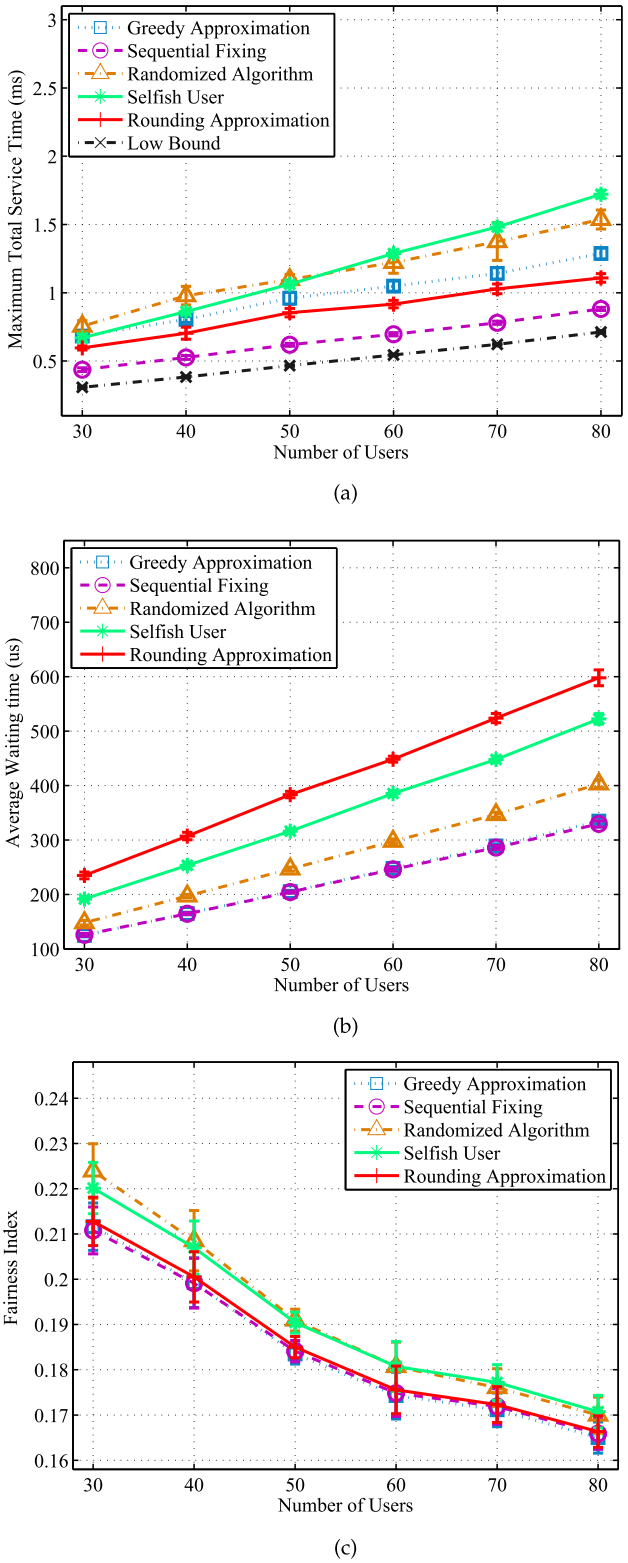


FIGURE 2. Performance evaluation of the open access strategy. (a) Total service time vs. number of users. (b) Average waiting time vs. number of users. (c) Fairness vs. number of users.

For comparison, we randomly schedule users in BS's in the selfish scheme and rounding approximation scheme. Intuitively, the larger the number of users, the larger the average

waiting time. We can see from the figure that, the average waiting time obtained by the greedy approximation algorithm is very close to that by the sequential fixing algorithm, while without appropriate scheduling, the rounding approximation algorithm achieves the largest waiting time, which is almost twice as large as the waiting time achieved by greedy approximation algorithm.

To evaluate the fairness performance, we adopt Jain's fairness index given by

$$\mathcal{J}(C_1, C_2, \ldots, C_N) = \frac{(\sum_{n=1}^{N} C_n)^2}{N \times \sum_{n=1}^{N} C_n^2},$$

where C_n is the throughput for user n [25]. The value of the index ranges from $1/N$ (worst case) to 1 (best case). It can be seen from Fig. 2(c) that fairness indexes decrease when the number of users is increased. We notice that, the selfish scheme and the randomized algorithm achieve better fairness than the other three schemes. Figs. 2(a) and 2(c) show that from operator's viewpoint, the selfish and the randomized schemes are not preferred since they produce less balanced load on BS's. From users's viewpoint, these two schemes may be appealing due to their fairness performance.

TABLE 3. Execution times of the proposed algorithms under the open access strategy (second).

No. of Users	30	40	50	60	70	80
Greedy Approx.	0.024	0.034	0.024	0.030	0.026	0.038
Sequential Fixing	16.532	24.020	30.809	48.713	47.842	50.654
Randomized Algorithm	0.030	0.048	0.077	0.136	0.132	0.151
Selfish User Scheme	0.035	0.035	0.035	0.035	0.036	0.026
Rounding Approx.	0.133	0.148	0.160	0.168	0.176	0.213

We list the execution times of the five schemes in Table 3. We find the execution time increases as the number of users is increased. The selfish scheme always has the smallest execution time, while sequential fixing has the largest execution time. Although the rounding approximation algorithm can achieve smaller load on the BS's, its execution time is greater than that of the greedy approximation algorithm. This result also justifies the complexity analysis for the proposed schemes. The execution time of the greedy approximation algorithm and the selfish scheme is always much smaller than other schemes and does not increase obviously with the number of users.

B. CLOSED ACCESS STRATEGY

We next investigate the second scenario with closed access femtocells. Now each FBS maintains a user list and only serves the listed users. Note that the MBS will always serve all the users inside its coverage.

In Fig. 3(a), we evaluate the impact of the number of users on total service time. Intuitively, the total service time increases as the number of users. However, we find that it also depends on the user list at each FBS. In the simulation, we randomly choose the user set $\mathcal{A}_m$ for BS m. Moreover, the user list at each FBS is further reduced due to the SINR

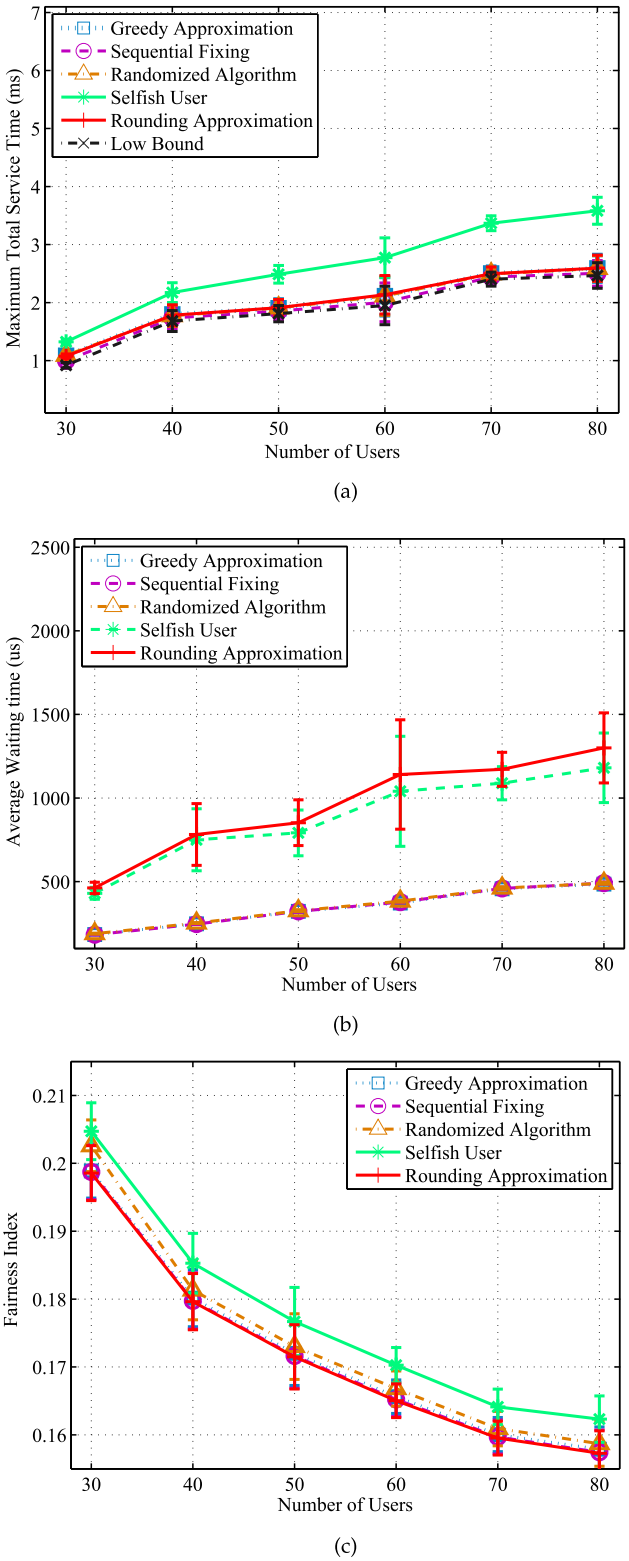


FIGURE 3. Performance evaluation of the closed access strategy. (a) Total service time vs. number of users. b) Average waiting time vs. number of users. (c) Fariness vs. number of users.

TABLE 4. Execution times of the proposed algorithms under the closed access strategy (second).

No. of Users	30	40	50	60	70	80
Greedy Approx.	0.003	0.004	0.005	0.005	0.007	0.006
Sequential Fixing	0.862	1.130	1.508	1.845	2.306	2.680
Randomized Algorithm	0.011	0.015	0.021	0.030	0.037	0.049
Selfish User Scheme	0.001	0.002	0.002	0.002	0.002	0.003
Rounding Approx.	0.029	0.028	0.032	0.034	0.037	0.040

threshold. Consequently, all the proposed algorithms achieve close performance in the closed access scenario. The total service time of the proposed algorithms is close to the low bound in closed access scenario. However, the performance of all the proposed algorithms is better than that of the selfish scheme, as we can see in Fig. 3(a).

We next show the impact of the number of users on average waiting time in Fig. 3(b). The scheduling policy setting is the same as that in the open access scenario. The result thus is also similar to the open access case that, the selfish scheme and the rounding approximation scheme achieve the largest waiting time. Actually with proposed optimal service scheduling, the approximation algorithms will achieve as less waiting time as that of the sequential fixing scheme.

We plot the fairness indices in Fig. 3(c). The randomized algorithm, although not better than the selfish scheme, achieves the best performance in fairness than the other proposed schemes. Despite of its good performance in minimizing the maximum service time, the rounding approximation algorithm, is not competitive with respect to fairness. Due to the randomness of user lists at BS's, the confidential intervals are larger than those in the open access scenario.

Finally, we list the execution time of five schemes in Table 4. As expected, the execution time increases as the number of users. Compared with Table 3, the closed access scheme requires less execution time than the open access scheme, which is due to the fact that there are more users allowed to get services in open access scheme.

VII. CONCLUSION

In this paper, we investigated the problem of cell association and service scheduling in two-tier femtocell networks. We developed several algorithms and analyzed their performance. The sequential fixing algorithm achieves the best performance in total service time but it has a relatively high complexity. Then we presented two approximation algorithms with lower complexity and proven approximation ratios. We also proposed a randomized algorithm with a proven performance bound that requires the least information exchange among users. In addition, we solved the service scheduling problem with an optimal solution. The proposed algorithms were validated with simulations in both open and closed access scenarios.

REFERENCES

[1] (Jun. 2014). "The 1000× mobile data challenge: More small cells, more spectrum, higher efficiency," Qualcomm, San Diego, CA, USA, Tech. Rep. [Online]. Available: http://www.qualcomm.com/solutions/wireless-networks/technologies/1000x-data

[2] I. K. Son, S. Mao, M. X. Gong, and Y. Li, "On frame-based scheduling for directional mmWave WPANs," in *Proc. IEEE INFOCOM*, Orlando, FL, USA, Mar. 2012, pp. 2149–2157.

[3] Z. He and S. Mao, "Adaptive multiple description coding and transmission of uncompressed video over 60 GHz networks," *ACM Mobile Comput. Commun. Rev.*, vol. 18, no. 1, pp. 14–24, Jan. 2014.

[4] I. K. Son, S. Mao, Y. Li, M. Chen, M. X. Gong, and T. S. Rappaport, "Frame-based medium access control for 5G wireless networks," *Mobile Netw. Appl.*, to be published, doi: 10.1007/s11036-014-0565-0.

[5] J. G. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. C. Reed, "Femtocells: Past, present, and future," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 497–508, Apr. 2012.

[6] J. Kleinberg and E. Tardos, *Algorithm Design*. Boston, MA, USA: Addison-Wesley, 2005.

[7] D. Hu, S. Mao, Y. T. Hou, and J. H. Reed, "Scalable video multicast in cognitive radio networks," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 3, pp. 334–344, Apr. 2010.

[8] G. de la Roche, A. Valcarce, D. Lopez-Perez, and J. Zhang, "Access control mechanisms for femtocells," *IEEE Commun. Mag.*, vol. 48, no. 1, pp. 33–39, Jan. 2010.

[9] D. Hu and S. Mao, "On medium grain scalable video streaming over femtocell cognitive radio networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 641–651, Apr. 2012.

[10] C. Dhahri and T. Ohtsuki, "Learning-based cell selection method for femtocell networks," in *Proc. IEEE 75th VTC Spring*, Yokohama, Japan, May 2012, pp. 1–5.

[11] R. Madan, J. Borran, A. Sampath, N. Bhushan, A. Khandekar, and T. Ji, "Cell association and interference coordination in heterogeneous LTE-A cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1479–1489, Dec. 2010.

[12] S. Corroy, L. Falconetti, and R. Mathar, "Dynamic cell association for downlink sum rate maximization in multi-cell heterogeneous networks," in *Proc. IEEE ICC*, Aachen, Germany, Jun. 2012, pp. 2457–2461.

[13] H. Zhou, D. Hu, S. Mao, P. Agrawal, and S. A. Reddy, "Cell association and handover management in femtocell networks," in *Proc. IEEE WCNC*, Shanghai, China, Apr. 2013, pp. 661–666.

[14] H.-S. Jo, Y. J. Sang, P. Xia, and J. G. Andrews, "Heterogeneous cellular networks with flexible cell association: A comprehensive downlink SINR analysis," *IEEE Trans. Wireless Commun.*, vol. 11, no. 10, pp. 3484–3495, Oct. 2012.

[15] S. Mukherjee, "Downlink SINR distribution in a heterogeneous cellular wireless network with biased cell association," in *Proc. IEEE ICC*, Ottawa, ON, Canada, Jun. 2012, pp. 6780–6786.

[16] Y. Xu and S. Mao. (Jan. 2015). "User association in massive MIMO HetNets." [Online]. Available: http://arxiv.org/abs/1501.03407

[17] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam, "Distributed α-optimal user association and cell load balancing in wireless networks," *IEEE/ACM Trans. Netw.*, vol. 20, no. 1, pp. 177–190, Feb. 2012.

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

[19] Y. Azar, J. Naor, and R. Rom, "The competitiveness of on-line assignments," in *Proc. 3rd Annu. ACM-SIAM Symp. Discrete Algorithms*, Orlando, FL, USA, Sep. 1992, pp. 203–210.

[20] A. Golaup, M. Mustapha, and L. B. Patanapongpibul, "Femtocell access control strategy in UMTS and LTE," *IEEE Commun. Mag.*, vol. 47, no. 9, pp. 117–123, Sep. 2009.

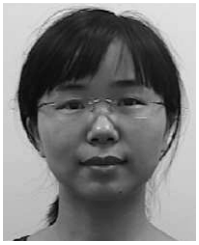
[21] S. F. Hasan, N. H. Siddique, and S. Chakraborty, "Femtocell versus WiFi—A survey and comparison of architecture and performance," in *Proc. 1st Int. Conf. Wireless VITAE*, Aalborg, Denmark, May 2009, pp. 916–920.

[22] Y. T. Hou, Y. Shi, and H. D. Sherali, "Spectrum sharing for multi-hop networking with cognitive radios," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 1, pp. 146–155, Jan. 2008.

[23] D. B. Shmoys and E. Tardos, "An approximation algorithm for the generalized assignment problem," *Math. Program.*, vol. 62, no. 3, pp. 461–474, Dec. 1993.

[24] J.-M. Moon and D.-H. Cho, "Novel handoff decision algorithm in hierarchical macro/femto-cell networks," in *Proc. IEEE WCNC*, Sydney, Australia, Apr. 2010, pp. 1–6.

[25] R. K. Jain, D.-M. Chiu, and W. R. Hawe. (Sep. 1984). "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems," DEC, Hudson, MA, USA, Tech. Rep. TR-301. [Online]. Available: http://www1.cse.wustl.edu/~jain/papers/ftp/fairness.pdf



HUI ZHOU received the Ph.D. degree in electrical and computer engineering from Auburn University, Auburn, AL, USA, in 2014, and the B.S. and M.S. degrees in electronic and information engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2007 and 2009, respectively. She was a Senior DSP Software Engineer with Zhongxing Telecommunication Equipment Corporation, Shanghai, China, from 2009 to 2011. She is currently a Software Development Engineer with Amazon, Seattle, WA, USA. Her research interests include femtocell networks and free space networks. She was a co-recipient of the 2013 IEEE International Conference on Communications Best Paper Award.

SHIWEN MAO (S'99–M'04–SM'09) received the Ph.D. degree in electrical and computer engineering from Polytechnic University, Brooklyn, NY, USA. He is currently the McWane Associate Professor with the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL, USA. His research interests include wireless networks and multimedia communications, with current focus on cognitive radio, small cells, millimeter-wave networks, free space optical networks, and smart grid. He was a Distinguished Lecturer of the IEEE Vehicular Technology Society—Class 2014. He is on the Editorial Board of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the IEEE INTERNET OF THINGS JOURNAL, and the IEEE COMMUNICATIONS SURVEYS AND TUTORIALS. He is the Vice Chair–Letters and Member Communications of the IEEE ComSoc Multimedia Communications Technical Committee. He was a recipient of the IEEE ComSoc MMTC Outstanding Leadership Award in 2013 and the NSF CAREER Award in 2010. He was a co-recipient of the 2013 IEEE ICC Best Paper Award and the 2004 IEEE Communications Society Leonard G. Abraham Prize in the field of communications systems.

PRATHIMA AGRAWAL (F'89) is currently the Samuel Ginn Distinguished Professor of Electrical and Computer Engineering and the Director of the Wireless Engineering Research and Education Center with Auburn University, Auburn, AL, USA. She was with Telcordia Technologies (formerly Bellcore), Morristown, NJ, USA, and AT&T/Lucent Bell Laboratories, Murray Hill, NJ, USA. She created and served as the Head of the Department of Networked Computing Research, Murray Hill. She is widely published and holds 51 U.S. patents. She received the B.E. and M.E. degrees in electrical communication engineering from the Indian Institute of Science, Bangalore, India, and the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 1977.