Femto-Matching: Efficient Traffic Offloading in Heterogeneous Cellular Networks

Wei Wang, Xiaobing Wu, Lei Xie and Sanglu Lu
State Key Laboratory for Novel Software Technology,
Department of Computer Science and Technology, Nanjing University, China
Email: {ww, wuxb, lxie, sanglu}@nju.edu.cn

***Abstract*—Heterogeneous cellular networks use small base stations, such as femtocells and WiFi APs, to offload traffic from macrocells. While network operators wish to globally balance the traffic, users may selfishly select the nearest base stations and make some base stations overcrowded. In this paper, we propose to use an auction-based algorithm – Femto-Matching, to achieve both load balancing among base stations and fairness among users. Femto-Matching optimally solves the global proportional fairness problem in polynomial time by transforming it into an equivalent matching problem. Furthermore, it can efficiently utilize the capacity of randomly deployed small cells. Our trace-driven simulations show Femto-Matching can reduce the load of macrocells by more than 30% compared to non-cooperative game based strategies.**

I. Introduction

In recent years, there is a trend for wireless cellular networks to incorporate different types of accessing technologies to meet the fast growing mobile traffic demands [1]. Unlike traditional cellular networks, where a single-tier of macrocells provides coverage over a large area, next generation wireless networks utilize multiple tiers of small Base Stations (BSs), including microcells, picocells, femtocells and WiFi APs, to offload mobile traffic from marcocells. In these Heterogeneous Networks (HetNets), a single mobile device can be covered by several BSs at the same time. For example, it is not uncommon to have more than five WiFi APs available to mobiles in urban areas [2]. How to select the right BS among all these nearby BSs for users becomes a critical issue for HetNets.

There are several challenges that are unique to the serving BS selection problem in HetNets. Firstly, users and network operators have misaligned objectives. Users wish to be associated with a BS which can provide the highest data throughput. However, decisions based on users' preference usually lead to imbalanced network traffic, i.e., some base stations are overcrowded while others remain idle. To improve network efficiency, network operators would like to globally balance the traffic. This may be unfair to some users as they are forced to be associated with a less preferred BS. Due to the mobility of users, these temporarily "bad" choices may lead to long-term benefits in average throughput. However, most existing association algorithms utilize local preferences over a static snapshot of the network topology [3], [4], which prevents the possibility of long-term traffic balancing.

Secondly, small cell base stations are randomly deployed. Unlike macrocells which are deployed with careful planning, small cell base stations, such as femtocell and WiFi APs, are often deployed by users in an ad-hoc manner. This leads to intrinsic spatial imbalance in network resource provisioning, i.e., certain areas in the network may have more small cells than others. To globally balance the traffic, users should be "pushed" towards regions with higher small cell density, so that their traffic gets a better chance to be offloaded by small cells. However, users and small cells only have limited local views on network topology. Therefore, it is hard to achieve global balance through distributed algorithms.

In this paper, we design an auction-based algorithm, called Femto-Matching, to address the above challenges. We show that we can achieve global optimality by carefully designing the auction mechanism. In our algorithm, the load of a BS is reflected by its price and users evaluate BSs based on how much improvement that the best BS can provide over the secondary choice. We prove that our design leads to an optimal solution in the sense of global proportional fairness.

One important observation gained from our design and analysis is that global optimization is crucial to the offloading efficiency for HetNets. With global matching schemes, it is possible to fully utilize the available resources provided by randomly deployed small BSs. In this way, the load of macrocells can be greatly reduced so that the network deployment cost is minimized. Our trace-driven simulations show that Femto-Matching can reduce the load of macrocells by more than 30% compared to non-cooperative game based strategies.

The main contributions of this work are as follows:

– We propose a new auction-based algorithm which can achieve the optimal solution for the proportional fair user association problem.

– We are the first to study offloading efficiency in random networks under different user association strategies. We prove that the ratio of users that cannot be offloaded by the optimal matching scheme is $O(\lambda_f^{-1/2}R^{-1})$ for homogenous Poisson Point Process, where λ_f is the density of femtocells and R is the communication range of femtocells.

II. Related Work

The load balancing problem in HetNets has been intensively studied in recent years [5]. One of the basic approaches to increase the number of users served by small cells is to introduce SINR Biasing, which encourages users to associate with a small cell even when the perceived SINR of the small cell is lower than the SINR of the macrocell [6], [7]. Global optimization algorithms are also proposed in [6], [8], [9], where the problem is formulated as a mixed integer programming and solved through linear relaxation or brute-force searching.

Unlike above global optimization approaches, game theory based algorithms utilize user preference to select the best BS in a distributed way. It is shown in [3] that the BS selection game converges to Nash equilibria when there is only one class of radio access technology. However, only carefully designed game strategies can converge in case there are multiple classes of BSs in the network. When both the preference of the user and the small cells are considered, a college admission algorithm can be used to find a stable matching between users and femtocells [4]. Unfortunately, these distributed algorithms do not provide any performance guarantee for randomly deployed networks.

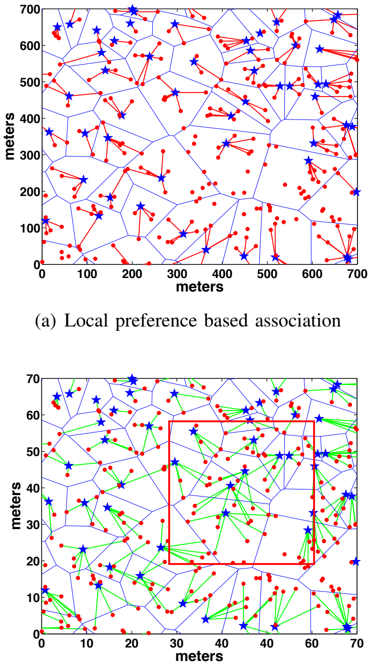
III. Motivation and System Model

A. Motivation

As a motivating example, Fig. 1 gives a snapshot of user association plan for randomly distributed users and BSs in a 70×70 meters area. To simplify the illustration, we only draw a single tier of femtocells, where each femtocell can serve up to 6 users. Fig. 1(a) shows the result when users select serving BS based on local preference, e.g., they try to associate with the BS which provides the highest throughput as in [3], [4]. In the lower-right region of Fig. 1(a), there is a cluster of users not associated with any femtocells, since all nearby femtocells do not have any vacancy. Although there are abundant lightly loaded femtocells in the upper-right region, they cannot serve these "orphan" users due to the limited communication range. This imbalanced traffic load for femtocells leads to higher traffic burdens on the next tier of BSs, i.e., these "orphan" users have to turn to the microcell tier or the marcocell tier for help.

Fig. 1(b) shows a more balanced association plan constructed through global optimization algorithm. In this case, lightly loaded femtocell can take over users associated with heavily loaded femtocells in a cascading way so that more users can be served by femtocells, as illustrated in the red rectangle of Fig. 1(b). Some of these femtocells serve users outside their Voronoi cells. This means the global optimization solution forces users to accept a less preferred choice by associating them to a femtocell which is not the nearest one to them. Existing association algorithms, such as SINR Biasing [7] or Game Theory based algorithm [3], cannot sacrifice the welfare of some users to achieve global optimality.

The difference between the user association plans in Fig. 1(a) and 1(b) leads to drastic difference in the load of macrocells. For example, femtocells serve about 80% users in Fig. 1(a). In a three tier HetNet which consists of femtocells, microcells and macrocells, the marcocell tier needs to serve $(1-80\%)\times(1-80\%) = 4\%$ users, assuming the microcell tier can also offload 80% users. If the offloading ratio is improved to 95% as in Fig. 1(b), the macrocell only needs to serve 0.25% of all users, which is an order of magnitude less than the previous case. In consequence, fewer costly macrocells need to be deployed or upgraded. Therefore, achieving an offloading ratio close to 100% is crucial for reducing the network cost.



(a) Local preference based association

(b) Globally optimized association

Fig. 1. User association for randomly deployed femtocells (femtocells: blue stars, users: red dots, association relationship: red or green lines).

B. System Model

Hinted by the observations in the previous section, we formulate the user association problem as a global optimization problem. Consider a HetNet which consists of N BSs in the BS set $\mathcal{B}$ and M users in the user set $\mathcal{U}$. Base Stations can be either macrocells, picocells, femtocells or WiFi APs. Different types of BSs are characterized by their Radio Access Technologies (RATs), transmission powers and backhaul bandwidths. We assume that users can switch between multiple BSs, but they can only associate with one BS at the same time, due to the capability limitations of mobile devices. If a user has multiple interfaces that can operate independently, we treat each interface as a separate user.

We assume that the achievable throughput for a given user is determined by two factors, the transmission bit rate and resources allocated to the user. Firstly, the wireless transmission bit rate for a given user is determined by the received SINR of the serving BS. As instantaneous SINR perceived by the user is time-varying due to slow/fast fading and transmission of nearby BSs/users, we use the long-term average rate to characterize the achievable rate. In the following discussions, we assume the average rate that user i can receive from BS j is r_{ij}. Secondly, the throughput for user i is also determined by the amount of wireless resources that BS j allocates for user i. By wireless resources, we mean time slots in TDMA systems, subcarriers in OFDM systems or Resource Blocks (RBs) in LTE. We use c_{ij} to denote the proportion of resources that BS j allocates to user i. Combining these two factors, the throughput of user i under BS j can be written as $c_{ij}r_{ij}$.

The goal of the user association problem for HetNet is to find an optimal association scheme for each user i along

TABLE I. NOTATIONS

Symbol	Description
r_{ij}	average transmission rate of user i under BS j
c_{ij}	resource allocated by BS j to user i
a_{ij}	association indicator for user i and BS j
n_j	the number of users covered by BS j
K_j	the number of users associated to BS j
κ	maximum number of users that femtocells can serve
λ_f	the density of femtocells
λ_u	the density of users
η	the ratio of users served by femtocells

with a resource allocation plan for each base station j so that the overall system utility is maximized. We choose the logarithm function as our utility function, i.e., user i has utility of $\log(x_i)$ when the throughput for user i is x_i. It is well known that logarithm utility functions lead to proportional fairness among users and it gives a balanced solution for both fairness and system efficiency [10]. In practice, Proportional Fairness Scheduling (PFS) is also a widely used scheduling algorithm for LTE eNBs [11].

We define the Proportional Fair User Association (PFUA) problem as:

$$\max \quad \textstyle\sum_i \log\left(\sum_j c_{ij} r_{ij} a_{ij}\right), \tag{1}$$

$$s.t. \quad \textstyle\sum_i c_{ij} \le 1 \quad \forall j \in \mathcal{B}, \tag{2}$$

$$\textstyle\sum_j a_{ij} \le 1 \quad \forall i \in \mathcal{U}, \tag{3}$$

$$a_{ij} \in \{0,1\},\, c_{ij} \ge 0 \quad \forall i \in \mathcal{U}, j \in \mathcal{B}. \tag{4}$$

In the above optimization problem, a_{ij} is the association indicator for user i and BS j. If user i is associated to BS j, we set $a_{ij} = 1$ and otherwise we set $a_{ij} = 0$. Therefore, $\sum_j c_{ij} r_{ij} a_{ij}$ in Eq. (1) is the total throughput that user i can get from the serving BS. Constraint (2) ensures that BS j will not over-utilize its resources. Constraint (3) ensures that each user can only be associated to one BS. The symbols used in this paper are summarized in Table I.

IV. ALGORITHM DESIGN

A. *Equivalent Matching Problem*

PFUA is a mixed integer programming problem where a_{ij} can only take integer values. Although a general extention of the PFUA – the GPF1 problem in [12] has been proven to be NP-hard, the hardness of PFUA is not known until recently. Most existing work on this topic uses two types of relaxations to solve this problem approximately. One way is to allow users to associate with multiple BSs at the same time, so that a_{ij} becomes a continuous variable [13], [6], [8] and the solution can be approximated via rounding the result of the relaxed convex optimization problem. The other way is to fix the number of users that each BS serves [12], [9]. In this way, the optimal throughput for each user-BS pair is also fixed, so that the problem can be directly reduced to a maximum weighted bipartite matching problem. The optimal solution can then be found by exhaustively searching over all possible combinations on the user number for each BS. While preparing the camera ready version of this paper, we found that a parallel work by Prasad et al. also proposed to use virtual BS based matching scheme to optimally solve the PFUA problem [14], [15]. Compared to their work, our auction based solution reveals structure of this problem and leads to a naturally distributed algorithm.

PFUA can be solved in polynomial time by casting the problem into a matching problem without fixing the number of users in each BS. We divide PFUA into two subproblems:

I.) Optimize resource allocation for a single BS

The goal of this subproblem is to find the optimal proportional fair share c_{ij} when there are K_j users associated to BS j. This problem can be easily solved as the optimal resource allocation has a closed form solution of $c_{ij} = 1/K_j$ [6], [3].

II.) Find a global optimal association scheme

The goal of this subproblem is to balance the number of users associated to each BS so that global utility can be maximized. This subproblem handles the dynamics in K_j for each BS, where larger K_j leads to a more crowded BS and lower throughput for each user. This subproblem is solved by constructing a bipartite graph using the solution of subproblem I and finding a maximum weighted matching on it.

Construction of the User Association Graph

Given user set $\mathcal{U}$ and BS set $\mathcal{B}$, we construct a bipartite graph $G = (U, V, E)$ as follows. We introduce *user* node u_i in the first vertex set U for every user $i \in \mathcal{U}$. We introduce n_j *virtual BS (VBS)* nodes $v_j^1, v_j^2, \ldots, v_j^{n_j}$ in the second vertex set V for every BS $j \in \mathcal{B}$, where n_j equals the number of users within communication range of BS j. We add an edge (u_i, v_j^k) between u_i and $v_j^k, k = 1, \ldots, n_j$, with weight of:

$$w_{ij}^k = \log\left(\frac{(k-1)^{k-1} r_{ij}}{k^k}\right),\ \forall\, k, \tag{5}$$

when user i is within the communication range of BS j.

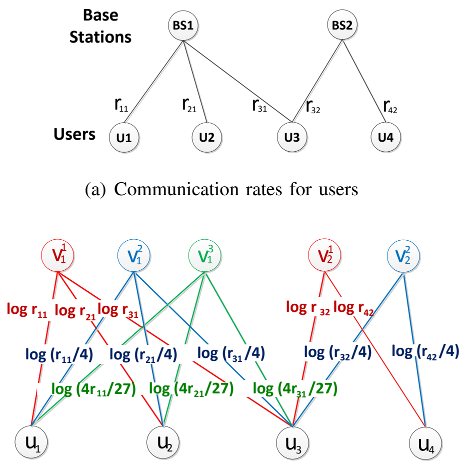
As an example, consider the network in Fig. 2(a), which has four users and two BSs with communication rates shown beside the corresponding links. Since there are three users U1, U2 and U3 within BS1's communication range, we split BS1 to three VBS nodes v_1^1, v_1^2 and v_1^3 in G. The VBS nodes are connected to user nodes with different weights to account for the dynamics in the number of users associated to the given BS.

An intuitive explanation of this construction is as follows. When the first user, say U1, is associated to BS1, the corresponding user node u_1 will be matched to v_1^1 with edge weight of $\log r_{11}$. This weight is equal to the utility for U1, as BS1 will allocate all its resources to U1. When a new user U2 is associated with BS1, the proportional fairness scheduler will equally divide the resource of BS1 between U1 and U2. Therefore, they get utility of $\log(r_{11}/2)$ and $\log(r_{21}/2)$, respectively. The marginal utility gain for adding U2 is:

$$\log(r_{11}/2) + \log(r_{21}/2) - \log r_{11} = \log(r_{21}/4), \tag{6}$$

which is equal to the edge weight between u_2 and v_1^2. By matching the kth user associated to BS1 to VBS node v_1^k, we actually keep track of the marginal utility gain that the user brings to BS1. In this way, we can handle the changing number of associated users in each BS and achieve global optimality in utility.

To prove the equivalence of these two problems, we first use the following lemma to show that maximum weighed matching on a subgraph of user association graph can find the



(a) Communication rates for users

(b) Constructed bipartite graph

Fig. 2. Bipartite graph construction for user association problem.

optimal solution for subproblem I: optimal resource allocation for a single BS.

Lemma 1: The maximum utility sum for BS j, when K_j users $j_1, j_2, \dots, j_{K_j}$ are associated to BS j is equal to the weight for maximum weighted matching on a subgraph $G' = (U', V', E')$ of G, induced by the node set $U' = \{u_{j_1}, u_{j_2}, \dots, u_{j_{K_j}}\}$, $V' = \{v_j^1, v_j^2, \dots, v_j^{n_j}\}$.

Proof: See Appendix A for the proof of Lemma 1. ■

Lemma 1 leads to the following Theorem:

Theorem 1: The maximum utility sum of the PFUA problem in Eq. (1)–(4) is equal to the weight of maximum weighted matching on the constructed user association graph G.

Proof: See Appendix B for the proof of Theorem 1. ■

Complexity Analysis

For a network with N Base Stations and M users, the computational cost for constructing the user association graph is $O(M^2N)$ and extracting the association plan from the matching result takes at most $O(M)$ time. There are $O(MN)$ VBS nodes and $O(M)$ user nodes in the user association graph. Thus, it takes $O(M^3N^3)$ time to solve the maximum weighted matching problem on the user association graph using Hungarian Algorithm [16]. Therefore, the overall time complexity for a centralized algorithm is $O(M^3N^3)$. Actually, this complexity bound is quite loose. As we will show in Sec.V, usually it is enough to include just $O(M)$ VBS nodes in the user association graph. The complexity can be reduced to $O(M^3)$ in this case.

B. Distributed Auction-based Algorithm

It is well known that maximum weighted matching problem can be solved through an auction algorithm in a distributed way [17]. However, directly applying a general solution leads to complex interactions between users and BSs. We observe that the weight w_{ij}^k in Eq. (5) consists of two parts $\log r_{ij}$ and $\log \frac{(k-1)^{k-1}}{k^k}$. The first part $\log r_{ij}$ is the rate for users and the second part $\log \frac{(k-1)^{k-1}}{k^k}$ is the penalty on the number of users associated with the BS. Therefore, these two parts can be maintained separately by users and BSs. Using this special structure of the problem, we design a distributed auction-based algorithm, called Femto-Matching, as follows:

i) Initialization

In the initialization phase, BSs generates VBS Nodes v_j^k with price $p_j^k = -\log \frac{(k-1)^{k-1}}{k^k}$. Every user i measures its rates to neighboring BSs and set the utility for associating with BS j as $c + \log r_{ij}$ where c is a predefined constant which is large enough to make all utilities larger than initial prices p_j^k. All VBSs and users are not assigned in this phase.

ii) Auction

The auction process is divided into multiple rounds. At the beginning of each round, BS j will announce $price_j$, which is the minimum price among all of its VBSs, to the neighboring users. Users will calculate the margin m_{ij}, which is $c + \log r_{ij} - price_j$. Each unassigned user find the BSs which provide the highest margin m_i^* and the second highest margin m_i'. User i submits bids of value $m_i^* - m_i'$ to the BS with the highest margin. In case of a tie, user i submits bid of a small value ε to one of the BSs. BSs pick the user with the highest bid as the winner and temporarily assign the winner to the VBS. The price of this VBS is raised by the amount of the highest bid and the user previously assigned to this VBS is unassigned. A new round of auction starts after BSs announce the new temporary assignment scheme. The price and temporary assignment announcement can use the broadcast mechanism provided by wireless networks, so that the number of interaction messages can be reduced.

iii) Termination

The auction process terminates when the assignment scheme stops changing. BSs will announce the final assignment scheme to the neighboring users.

For example, consider the network in Fig. 2. Suppose that we have $r_{11} = r_{31} = 3$, $r_{21} = r_{32} = r_{42} = 2$ and $c = 2$. At the initial state, prices for VBS nodes will be set as the first row in Table II. Both BS1 and BS2 will announce price of 0 in the first round. U1's highest margin m_1^* is $2 + \log 3$ (associating with v_1^1) and there is no secondary choice for U1. So, U1 will submit bid of $2 + \log 3$ to BS1. Similarly, U2 and U4 will submit bid of $2 + \log 2$ to BS1 and BS2, respectively. U3 has two choices and BS1 gives higher margin. So, U3 will submit bid of $2 + \log 3 - (2 + \log 2) = \log 3/2$ to BS1. In this round, U1 wins the VBS v_1^1, U4 wins the VBS v_2^1 and the prices of these VBSs are raised. After the price adjustment, both BS1 and BS2 announce price of $\log 4$ (for v_1^2 and v_2^2) in the second round. In this round, U2 and U3 will both bid for v_1^2. As the bid submitted by U2 is $2 - \log 2$ which is larger than the bid of $\log 3/2$ submitted by U3, U2 wins irrespective of its lower transmission rate compared to U3. In the third round, U3 will compare the margin provided by v_1^3 and v_2^2 and choose to associate with BS2. We can verify that this solution actually maximizes the utility sum in PFUA.

Detailed algorithm for Femto-Matching is shown in Algorithm 1 and 2. Our algorithm uses the Jacobi version of

TABLE II. PRICE FOR VBS IN THE AUCTION SAMPLE

rounds	p_1^1	p_1^2	p_1^3	p_2^1	p_2^2
0	**0**	$\log 4$	$\log(27/4)$	**0**	$\log 4$
1	$2+\log 3$	**log 4**	$\log(27/4)$	2 + log 2	**log 4**
2	$2+\log 3$	2 + log 2	**log (27/4)**	2 + log 2	**log 4**
3	$2+\log 3$	2 + log 2	**log (27/4)**	2 + log 2	**log (9/2)**

Algorithm 1 Auction procedure for BS j

```
$p_j^k \leftarrow -\log \frac{(k-1)^{k-1}}{k^k}$, $\forall\, k$.
while Assignment changes do
    $price_j \leftarrow \min_k\{p_j^k\}$, announce $price_j$
    $k^* \leftarrow arg\, min_k\{p_j^k\}$
    Collect $bid_i$ from neighboring users
    $winning\_user \leftarrow arg\, max_i\{bid_i\}$
    Temporarily assign VBS with index $k^*$ to the $winning\_user$ and remove the user previously assigned with VBS $k^*$
    $p_j^{k^*} \leftarrow p_j^{k^*} + bid_{winning\_user}$
    Announce the new temporary assignment
end while
Announce the final assignment
```

the auction where all unassigned users submit their bids in the same round. It is also possible to use the Gauss-Seidel version of auction, which allows users to submit bids in an asynchronous manner.

As Femto-Matching essentially solves the dual problem of the weighted bipartite matching problem, we can show that the solution is within $M\varepsilon$ to the optimal solution in a similar way as in [17]. Note that the price for the announced VBS increases by at least ε in each round, therefore the maximum number of rounds for biding is upper bounded by $\max\{c+\log r_{ij}\} \times \kappa/\varepsilon$, where κ is the maximum number of VBS nodes that a BS can have. By tuning the value of ε, we can tradeoff between the convergence time and the approximation ratio.

C. Practical Issues

1) Handling user mobility

One advantage of Femto-Matching is that it can work in dynamical environments where users keep moving around. Once the assignment is calculated for a network snapshot, we can extend the algorithm to handle network dynamics as follows:

– When a new user joins the network, he can use Algorithm 2 to bid for a VBS under the current set of prices. The calculation of the new association plan only involves nearby BSs and users. In case that there is a vacant VBS in the nearby BS, the new user will be served by the vacant VBS. Otherwise, the new user may "kickout" one of the existing users and causes a cascaded handoff which involves multiple users. There are two ways to reduce the impact of cascaded handoff. The first one is to balance the vacant VBS by increasing the price of the last free VBS of each BS. Therefore, the last free VBS is always reserved for new comers. The other way is to introduce handoff penalty by increasing the price of VBS which is assigned to existing users.

– When a user leaves the network, the BS reduces the price of the VBS associated to that user to the initial value and asks neighboring users to bid for that VBS. This may also lead to cascaded reassignments. We can reduce the impact of this by asking users to be conservative in reassignment, e.g., requesting new bids to be larger than a given value.

Algorithm 2 Auction procedure for user i

```
Calculate $r_{ij}$ for each neighboring BS
while Assignment not finalized do
    Collect $price_j$ and temporary assignment from neighboring BSs
    if not temporarily assigned then
        $m_{ij} \leftarrow c + \log r_{ij} - price_j$, $\forall j$
        $j^* \leftarrow arg\, max_j\{m_{ij}\}$, $m_i^* \leftarrow max_j\{m_{ij}\}$
        $m_i' \leftarrow$ second largest value in $m_{ij}$
        if $m_i^* - m_i' > 0$ then
            Submit $bid_i = m_i^* - m_i'$ to BS $j^*$
        else
            Submit $bid_i = \varepsilon$ to BS $j^*$
        end if
    end if
end while
```

– When the device moves around and its communication rate changes, we can treat this case as a simultaneous user leave and join to get the new association plan.

It is possible to tune the price to reduce the number of handoffs when user mobility patterns need to be considered. For example, we can reduce the constant c for a moving user so that handoff happens only when transmission rate of the previous serving BS is lower than a given bound.

2) Multiuser diversity gain

Proportional Fairness Scheduling (PFS) implemented in 3G/4G networks can opportunistically schedule a user with better channel quality to improve the average throughput. As users experience independent fading and noise conditions, there is a multiuser diversity gain which can be achieved via PFS. For Rayleigh Fading channel, the average throughput for user i is given as $r_{ij}/K_j \times \sum_{k=1}^{K_j} \frac{1}{k}$, when there are K_j users under a PFS based BS [18]. By defining the multiuser diversity gain function as $g(K_j) = \sum_{k=1}^{K_j} \frac{1}{k}$, we can set the weight for edges connect u_i to VBS v_j^k as:

$$w_{ij}^k = \log\left(\frac{(k-1)^{k-1} g(k) r_{ij}}{k^k g(k-1)}\right). \tag{7}$$

By a similar procedure as in Sec. IV-A, we can show that our matching algorithm can also achieve the maximum utility for PFUA with this new weight function. As we can adjust the weights for each Base Station, our algorithm can work in HetNets which consist of both PFS based BS and non-PFS based BS at the same time.

3) Including multiple tiers of BSs

Our solution for PFUA can work for networks with different types of BSs. Small BSs such as femtocells may have limited service capability so that only a limited number of users can be associated to them. For these small BSs, we can impose a bound on the number of VBS Nodes for the given small BS. On the other hand, macrocells have significantly more resources and higher transmission powers than small BSs. We adjust the rate r_{ij} for the given macrocell j by multiplying the rate calculated through SINR with a modification factor b_j to reflect that the marcocell j has more wireless resources than small BSs.

V. Offloading Efficiency Analysis

In this section, we study the efficiency of association algorithms in a randomly deployed network to get a better understanding about the performance of matching based solutions. We mainly focus on the metric of offloading efficiency η, which is defined as the ratio of users which can be served by the femtocell tier under the given association scheme. In a network with multiple tiers, offloading efficiency determines how many users should be served by the higher tiers of BSs, as discussed in Sec. III.

We assume that both femtocells and the users are distributed as homogenous Poisson Point Process (PPP) with intensity of λ_f and λ_u, respectively. We define the load factor l as λ_u/λ_f, i.e., the average number of users that a femtocell should serve. We further assume that each femtocell can serve at most κ users within its communication range of R. Under these capacity and communication range constraints, we can analytically compare the efficiency of different association algorithms.

First, consider the association scheme which tries to associate users to the nearest femtocell. If the nearest femtocell is full, the user is associated to the higher tier of cells.

Define $\mathbb{P}\{N_v = k\}$ as the probability that the Voronoi cell of a femtocell contains k users. We have the offloading efficiency of the associate-to-nearest algorithm as:

$$\begin{aligned}\eta_n &= \frac{1}{l}\left(\sum_{k=0}^{\kappa} k\mathbb{P}\{N_v = k\} + \kappa \sum_{k=\kappa+1}^{\infty} \mathbb{P}\{N_v = k\}\right) \\ &= \frac{1}{l}\left(\kappa - \sum_{k=0}^{\kappa}(\kappa - k)\mathbb{P}\{N_v = k\}\right). \qquad (8)\end{aligned}$$

Unfortunately, there is no closed form solution for $\mathbb{P}\{N_v = k\}$, the best known approximation is given by [19], [20]:

$$\mathbb{P}\{N_v = k\} = \frac{3.5^{3.5}\Gamma(k+3.5)l^k}{\Gamma(3.5)k!(l+3.5)^{k+3.5}}, \qquad (9)$$

where $\Gamma(x)$ is the Gamma function.

We observe that the offloading efficiency η_n is only related to κ and l. Thus, it cannot be improved by increasing the network density while keeping κ and l fixed. This implies that the offloading efficiency for the associate-to-nearest algorithm is governed by the intrinsic randomness in node distribution. Table III gives numerical results for η_n. When $\kappa = l$, i.e., femtocells have just enough resources to serve all users, the associate-to-nearest algorithm have offloading efficiency lower than 75%. To achieve offloading efficiency higher than 95%, we often need $\kappa > 2l$, which means the capacity of femtocells should be over-provisioned by two times than the actual number of users to be served.

Algorithms using local preference can improve over the naive associate-to-nearest algorithm by associating users to nearby vacant femtocells. However, global matching schemes such as Femto-Matching can achieve the optimal offloading efficiency. This is because offload efficiency is maximized when the size of matching between users and femtocells is maximized. By using a matching algorithm considering the global topology, we can fully utilize the capability of femtocells, as shown by Theorem 2.

TABLE III. Offloading efficiency for the associate-to-nearest algorithm

l	$\kappa=1$	$\kappa=2$	$\kappa=3$	$\kappa=4$	$\kappa=5$	$\kappa=6$
1	**0.5851**	0.8474	0.9483	0.9835	0.9950	0.9985
2		**0.6636**	0.8230	0.9110	0.9568	0.9796
3			**0.6980**	0.8132	0.8877	0.9341
4				**0.7176**	0.8080	0.8721
5					**0.7303**	0.8048
6						**0.7393**

Theorem 2: The offloading efficiency of global matching based association algorithm is lower bounded by:

$$\eta_m = 1 - \sqrt{\frac{(1+l)\log 2}{\pi l \lambda_f R^2}}, \qquad (10)$$

with high probability, when both users and femtocell are distributed according to PPP with intensity of λ_u and λ_f with $\kappa \geq l$.

Proof: See Appendix C for the proof of Theorem 2. ■

Theorem 2 shows that the offloading efficiency quickly approaches 1 when the network density increases with a fixed value of κ and l. This hints that global matching algorithms could be a powerful way to smooth out the randomness in user/femtocell distributions. The actual performance comparison between global matching algorithm and local preference based algorithms is studied via simulations in Sec. VI.

VI. Simulation Results

A. Simulation Setup

Our simulation is conducted in randomly deployed networks within a 100 × 100 meters region. We assume that the marcocell is located at the center of the network while users and femtocells are distributed according to PPP. The communication rate for a user i under BS j is set to:

$$r_{ij} = \log\left(1 + \frac{P_j}{N_0 d_{ij}^{\alpha}}\right), \qquad (11)$$

where d_{ij} is the distance between user and BS. The meaning of other parameters and the default simulation setting is summarized in Tab. IV.

We compare our Femto-Matching algorithm with three association algorithms:

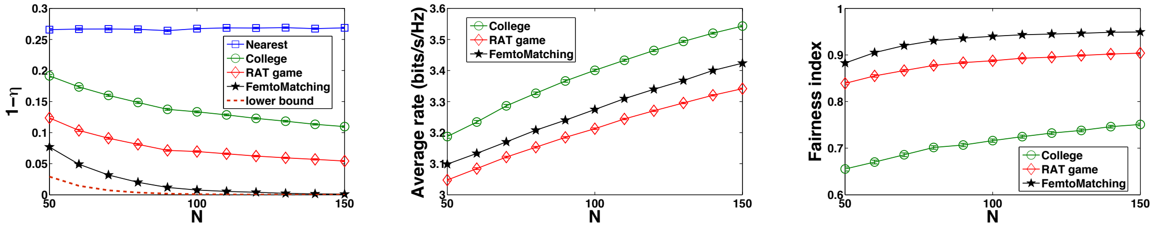
– *Associate-to-nearest*

This algorithm associates users to the nearest femtocell, as described in Sec.V.

– *RAT selection game* [3]

In this algorithm, users use their expected throughput as their preference. Users always try to switch to a BS which provides higher expected throughput.

– *College admission* [4]



(a) Ratio of users that cannot be offloaded

(b) Average rate

(c) Fairness among users

Fig. 3. Simulation result for different association algorithms in a 100×100 meters region, $\kappa = l = 5$ (95% confidence interval).

TABLE IV. SIMULATION PARAMETERS

Symbol	Description	Value
L	Simulation region size	100 meters
R	Transmission range	15 meters
α	Pathloss exponent	3
P_j	Transmission power	40 (macro), 20 (femto) dBm
N_0	Noise level	-90 dBm

In college admission algorithm, users rank nearby BS based on their preferences and submit requests to femtocells in sequence. Femtocells rank received requests and turn down lower ranked requests when the number of requests exceeds its capacity. In our implementation, we use average communication rate to determine the preference for both users and femtocells.

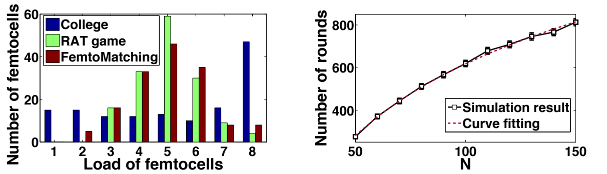
B. Performance Evaluation

Fig. 3 compares the performance of Femto-Matching with other association algorithms. As the number of femtocells in the simulation region increases, the ratio of users that cannot be offloaded, $1-\eta$, remains the same in the associate-to-nearest algorithm. On the other hand, Femto-matching has the highest decreasing speed in $1-\eta$ among all association algorithms. In most cases, the proportion of users that cannot be offloaded is less than half of the RAT selection game solution and it is very close to the lower bound, the ratio of users that has no femtocells within communication range.

The result of average throughput is given in Fig. 3(b). We see that college admission algorithm has the highest average throughput, as this algorithm tries to assign users to their nearest femtocell and users close to femtocells get very high rates. Femto-Matching outperforms RAT game in average throughput. This is because a large number of users are assigned to the marcocell in the RAT game algorithm, which makes the macrocell overcrowded and lowers the throughput for those associated with the macrocell. Fig. 3(c) shows Jain's fairness index for user throughput, which is defined as:

$$J = \frac{(\sum_{i=1}^{M} x_i)^2}{M \sum_{i=1}^{M} x_i^2}, \tag{12}$$

where $x_i = \sum_j c_{ij} r_{ij} a_{ij}$ is the throughput for user i. We see that Femto-Matching has the highest fairness index among all algorithms. The reason that college admission algorithm is poor in fairness is that users at borders of femtocell are often assigned to the macrocell, which gives them a much smaller throughput share compared to those associated to a nearby femtocell. Compared to college admission algorithm, Fetmo-Matching provides reasonable tradeoff by achieving much better fairness with a small reduction in average throughput.



(a) Load distribution for femtocells

(b) Number of rounds for Femto-Matching (95% confidence interval)

Fig. 4. Simulation result for randomly deployed networks with $\kappa = 8$, $l = 5$.

Fig. 4(a) shows the load distribution for femtocells when the capacity of femtocells $\kappa = 8$ is larger than the load $l = 5$ for a network with 150 femtcocells. We see that college admission algorithm gives unbalanced results, where more than 40 femtocells reach their capacity limit of 8 users. Femto-Matching achieves much better load balancing where about 1/3 femtocells are serving 5 users, which is exactly equal to the average load.

The number of auction rounds that Femto-Matching used for different network sizes is shown in Fig. 4(b). For networks with 150 femtocells and 750 users, Femto-Matching need only about 800 rounds of auctions. Using curve fitting, we find the number of auction rounds required under our simulation settings is $485.4 \log N - 1617.7$. Therefore, our simulation result hints that the number of rounds increase as $O(\log N)$.

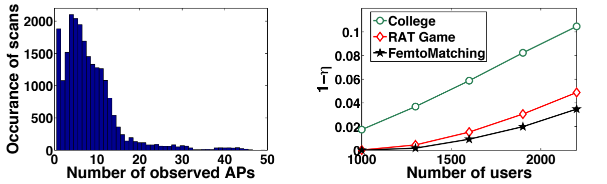
C. Trace-Driven Simulation

We use WiFi trace collected by the UIUC UIM system [21] to verify the usefulness of Femto-Matching in real networks. This database contains more than 22,000 WiFi scan records for 28 mobile phones during a period of 3 weeks. Fig. 5(a) is the histogram of the number of APs observed per-scan, where the average number of APs observed per-scan is 8.39.

In our trace driven experiment, we treat each unique scan as a virtual user and randomly pick M virtual users from all the scans as users to be offloaded. Fig. 5(b) compares the number of users cannot be offloaded when APs can serve up to $\kappa = 4$ users. Both RAT game and Femto-Matching outperform the college admission algorithm in real traces. Femto-Matching can further reduce the number of unmatched users by 30% compared to RAT game.

VII. CONCLUSION

In this paper, we proposed a new auction-based user association algorithm which uses matching between users and femtocells to improve the offloading efficiency of randomly



(a) Number of observed APs per-scan (b) Ratio of users that cannot be offloaded

Fig. 5. Trace driven simulation result.

deployed HetNets. There are several interesting issues which may deserve further study: The first issue is how does user mobility affect the performance of user association algorithm. The second is related to our assumption that users are cooperative and give truthful bids. In practical systems, how to design a mechanism to prevent malicious users from benefiting through the auction system needs to be considered.

ACKNOWLEDGEMENT

This work is partially supported by NSFC Grants (No.61373129, No.91218302, No.61373130, No.61100196, No.61472185 and No.61321491), key project of Jiangsu Research Program Grant (No.BE2013116), NSF of Jiangsu Province Grant (No.BK20141319), EU FP7 IRSES MobileCloud Project Grant (No.612212) and EU FP7 CROWN project (PIRSES-GA-2013-610524).

APPENDIX

A. Proof of Lemma 1

Proof: As the optimal c_{ij} is equal to $1/K_j$ [6], [3], the maximum utility sum for the K_j users is given by:

$$\sum_{k=1}^{K_j} \log\left(\frac{r_{j_k j}}{K_j}\right) = \log\left(\frac{\prod_{k=1}^{K_j} r_{j_k j}}{K_j^{K_j}}\right). \tag{13}$$

Now consider the maximum weighted matching in G'.

i) We first show that the weight for maximum weighted matching in G' is lower bounded by the utility given by Eq. (13). We can construct a matching in G' where user node u_{j_k} is matched to v_j^k. The sum of the weights for edges in this matching is given by:

$$\sum_{k=1}^{K_j} \log\left(\frac{(k-1)^{k-1} r_{j_k j}}{k^k}\right) = \log\left(\frac{\prod_{k=1}^{K_j} r_{j_k j}}{K_j^{K_j}}\right), \tag{14}$$

which is equal to the result in Eq.(13).

ii) We next show that the utility given by Eq.(13) is also an upper bound for the maximum weighted matching in G'. Consider the dual problem of maximum weighted matching, which assigns a non-negative price on each node and find the minimum price vertex cover in G' [16]. As edges in G' may have negative weights, we add two positive constants, c_1 and c_2, to the weight of each edge in G', i.e., $w'^k_{ij} = w^k_{ij} + c_1 + c_2$, to avoid negative weights. We set

$$c_1 = q(K_j), c_2 = |\min_i\{\log r_{ij}\}| + q(n_j),$$

where $q(k) = -\log\left(\frac{(k-1)^{k-1}}{k^k}\right)$. As $x \log x$ is an increasing convex function when $x \geq 1$, we can verify that $q(k) \geq 0$ and $q(k) < q(k+1)$.

With the above adjustments, we have weights $w'^k_{ij} \geq 0$ for all edges in G'. We then assign a price $p\{\cdot\}$ to each node in G', with:

$$p\{v_j^k\} = \begin{cases} c_1 & \text{if } k = 1 \\ c_1 - q(k) & \text{if } 1 < k \leq K_j, \\ 0 & \text{if } k > K_j \end{cases} \tag{15}$$

$$p\{u_i\} = c_2 + \log r_{ij} \qquad \forall i \in U'. \tag{16}$$

With this construction, we can verify that:

$$p\{v_j^k\} \geq 0, p\{u_i\} \geq 0, w'^k_{ij} \leq p\{u_i\} + p\{v_j^k\}, \quad \forall i \in U', \forall k.$$

For any matching $\mathcal{M}$ on G' we have:

$$\begin{aligned} \sum_{(u_i, v_j^k) \in \mathcal{M}} w'^k_{ij} &\leq \sum_{(u_i, v_j^k) \in \mathcal{M}} \Big(p\{u_i\} + p\{v_j^k\}\Big) \\ &\leq \sum_{u_i \in U'} p\{u_i\} + \sum_{v_j^k \in V'} p\{v_j^k\} \\ &= K_j c_2 + \sum_{k=1}^{K_j} \log r_{j_k j} + K_j c_1 + \sum_{k=2}^{K_j} \log\left(\frac{(k-1)^{k-1}}{k^k}\right) \\ &= K_j(c_1 + c_2) + \log\left(\frac{\prod_{k=1}^{K_j} r_{j_k j}}{K_j^{K_j}}\right). \end{aligned} \tag{17}$$

As G' is a complete bipartite graph with $|U'| = K_j \leq n_j = |V'|$, we always have K_j edges in the maximum weighted matching. By removing the weight adjustment factor on edges, which is equal to $K_j(c_1+c_2)$, we can see that $\sum_{(u_i, v_j^k) \in \mathcal{M}} w^k_{ij}$ for any matching on G' is upper bounded by the right side of Eq. (13). As the upper bound and lower bound for the maximum weighted matching are the same, the conclusion follows. ■

B. Proof of Theorem 1

Proof: We prove the equivalence of the optimal solution for PFUA and the maximum weighted matching problem by showing that their optimal solutions can be converted to each other and optimal values of the solutions are equal.

First, consider the optimal solution of the PFUA problem. Given the optimal user association indicator a_{ij}, we can partition the bipartite graph G into a set of subgraphs G_j, with each G_j only contains VBS nodes v_j^k of BS j and user nodes that are associated to BS j in the optimal solution of PFUA. All edges connecting VBS nodes and user nodes belonging to different subgraphs are removed. By Lemma 1, the weight for maximum weighted matching on each subgraph G_j is equal to the utility sum for each BS j in PFUA. Therefore, the matching weight sum over all G_j is equal to the maximum utility sum of the optimal solution for PFUA. As $\cup_j G_j \subseteq G$, maximum weighted matching in G is no less than the sum of maximum weighted matching on each subgraph. Therefore, the weight of maximum weighted matching in G is larger than or equal to the maximum utility sum of the PFUA problem.

Second, consider the maximum weighted matching on G. When the matching is given, we can generate an associate scheme, where user i is associated to BS j if u_i is matched to one of the VBS node v_j^k. In a similar way, we can also partition G into a set of subgraphs G_j. In this case, none of the edges in the maximum weighted matching will be deleted, so the maximum weighted matching on $\cup_j G_j$ is equal to the maximum weighted matching on G. By Lemma 1, there exists a scheme for PFUA that has the sum utility equal to the maximum matching weight for each G_j. As nodes in G_j are disjoint, we can construct a solution for PFUA with the utility sum equal to the maximum weighted matching on G. Combining the result of the two steps, the optimal value for the two problems must be equal to each other. ■

C. Proof of Theorem 2

Proof: We first consider the case where $\kappa = l$. Suppose there are more than $(1-\eta_m)M$ users cannot be matched to femtocells. Under this condition, with arguments similar to Hall's marriage Theorem, we can always find a subset of BSs $B' \subseteq \mathcal{B}$ where its neighborhood size:

$$\mathcal{N}(B') \le l|B'| - (1-\eta_m)M, \tag{18}$$

where $|B'|$ is the number of BSs in set B' and $\mathcal{N}(B')$ is the number of users within the communication range of BSs in B'. Note that we only need to consider subsets with size $|B'|$ larger than $(1-\eta_m)|\mathcal{B}|$, since otherwise the right side of Eq. (18) is smaller than 0.

When BSs are distributed as Poisson Point Process, the chance that a user is within the neighborhood of the subset B' of the BSs is given by [22]:

$$p_{B'} = 1 - e^{-\frac{\pi|B'|\lambda_f R^2}{|\mathcal{B}|}}. \tag{19}$$

As users are also distributed as PPP, $\mathcal{N}(B')$ follows binomial distribution with parameters of M and $p_{B'}$. By Chernoff bound, we have:

$$\mathbb{P}\left\{\mathcal{N}(B') \le l|B'| - (1-\eta_m)M\right\} \le e^{-D(\frac{l|B'|}{M}-(1-\eta_m)||p_{B'})M},$$

with: $$D\left(\frac{l|B'|}{M} - (1-\eta_m)||p_{B'}\right)$$
$$= \left(\frac{l|B'|}{M} - (1-\eta_m)\right)\log\frac{\frac{l|B'|}{M}-(1-\eta_m)}{p_{B'}}$$
$$+ \left(1 - \frac{l|B'|}{M} + (1-\eta_m)\right)\log\frac{1-\frac{l|B'|}{M}+(1-\eta_m)}{1-p_{B'}}$$
$$= -H\left(\frac{l|B'|}{M} - (1-\eta_m)\right) + \left(\frac{l|B'|}{M} - (1-\eta_m)\right)\log\frac{1}{p_{B'}}$$
$$+ \left(1 - \frac{l|B'|}{M} + (1-\eta_m)\right)\log\frac{1}{1-p_{B'}}, \tag{20}$$

where $H(x)$ is the entropy of x (in nats). Using the fact that $-H(x) \ge -\log 2$ and $(1-\eta_m) \le \frac{l|B'|}{M} \le 1$, we have:

$$D\left(\frac{l|B'|}{M} - (1-\eta_m)||p_{B'}\right) > (1-\eta_m)\log\frac{1}{1-p_{B'}} - \log 2.$$

The strict larger than sign comes from the fact that the first two terms in Eq. (20) cannot reach their minimum value at the same time. Using Eq. (10) and Eq. (19), we have:

$$D\left(\frac{l|B'|}{M} - (1-\eta_m)||p_{B'}\right) > \frac{\log 2}{l}, \tag{21}$$

as $\frac{|B'|}{|\mathcal{B}|} \ge 1 - \eta_m$.

Since $M = lN$, we get

$$\mathbb{P}\left\{\mathcal{N}(B') \le l|B'| - (1-\eta_m)M\right\} < 2^{-N(1+\epsilon)}, \tag{22}$$

where ϵ is a small constant that can be derived from Eq. (21). There are at most 2^N choices for B', since the number of femtocells is N. Using the union bound, the chance that there exists a subset B' with $\mathcal{N}(B') \le l|B'| - (1-\eta_m)M$ is $o(1)$. Consequently, the number of unmatched users is smaller than $(1-\eta_m)M$ with high probability when $N \to \infty$. When $\kappa > l$, the number of matched users is always larger than the case $\kappa = l$. Therefore, the conclusion of Theorem 2 follows. ■

References

[1] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update, 2013–2018," Cisco White Paper , 2014.

[2] H. Soroush, N. Banerjee, A. Balasubramanian, M. D. Corner, B. N. Levine, and B. Lynn, "Dome: a diverse outdoor mobile testbed," in *Proceedings of ACM HotPlanet*, 2009.

[3] E. Aryafar, A. Keshavarz-Haddad, M. Wang, and M. Chiang, "RAT selection games in hetnets," in *Proceedings of IEEE INFOCOM*, 2013.

[4] W. Saad, Z. Han, R. Zheng, M. Debbah, and H. V. Poor, "A college admissions game for uplink user association in wireless small cell networks," in *Proceedings of IEEE INFOCOM*, 2014.

[5] J. G. Andrews, S. Singh, Q. Ye, X. Lin, and H. S. Dhillon, "An overview of load balancing in HetNets: old myths and open problems," *IEEE Wireless Communications*, vol. 21, no. 2, pp. 18–25, 2014.

[6] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. Andrews, "User association for load balancing in heterogeneous cellular networks," *Wireless Communications, IEEE Transactions on*, vol. 12, no. 6, pp. 2706–2716, 2013.

[7] S. Singh and J. Andrews, "Joint resource partitioning and offloading in heterogeneous cellular networks," *IEEE Trans. Wireless Communications*, vol. 13, no. 2, pp. 888–901, 2014.

[8] W. Zhao, S. Wang, C. Wang, and X. Wu, "Cell planning for heterogeneous networks: An approximation algorithm," in *Proceedings of IEEE INFOCOM*, 2014.

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

[10] F. Kelly, "Charging and rate control for elastic traffic," *European transactions on Telecommunications*, vol. 8, no. 1, pp. 33–37, 1997.

[11] F. Khan, *LTE for 4G Mobile Broadband: Air Interface Technologies and Performance*. Cambridge University Press, 2009.

[12] T. Bu, L. Li, and R. Ramjee, "Generalized proportional fair scheduling in third generation wireless data networks," in *Proceedings of IEEE INFOCOM*, 2006.

[13] L. Li, M. Pal, and Y. R. Yang, "Proportional fairness in multi-rate wireless LANs," in *Proceedings of IEEE INFOCOM*, 2008.

[14] N. Prasad, M. Arslan, and S. Rangarajan, "Exploiting cell dormancy and load balancing in LTE HetNets: Optimizing the proportional fairness utility," in *IEEE ICC*, 2014.

[15] ——, "Exploiting cell dormancy and load balancing in LTE HetNets: Optimizing the proportional fairness utility," *IEEE Trans. Communications*, vol. 62, no. 10, pp. 3706 – 3722, 2014.

[16] R. E. Burkard, M. Dell'Amico, and S. Martello, *Assignment Problems*. Siam, 2009.

[17] D. P. Bertsekas, "A new algorithm for the assignment problem," *Mathematical Programming*, vol. 21, no. 1, pp. 152–171, 1981.

[18] E. Liu, Q. Zhang, and K. K. Leung, "Asymptotic analysis of proportionally fair scheduling in rayleigh fading," *IEEE Trans. Wireless Communications*, vol. 10, no. 6, pp. 1764–1775, 2011.

[19] J.-S. Ferenc and Z. Néda, "On the size distribution of poisson voronoi cells," *Physica A: Statistical Mechanics and its Applications*, vol. 385, no. 2, pp. 518–526, 2007.

[20] S. M. Yu and S.-L. Kim, "Downlink capacity and base station density in cellular networks," in *Proceddings of IEEE WiOpt*, 2013, pp. 119–124.

[21] K. Nahrstedt and L. Vu, "CRAWDAD data set uiuc/uim (v. 2012-01-24)," Aviable online, http://crawdad.org/uiuc/uim/, Jan. 2012.

[22] P. Hall, *Introduction to the theory of coverage processes*. John Wiley & Sons, Inc, 1988.