Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2017) 25: 1976 – 1992

doi:10.3906/elk-1510-69

Mobility and load aware radio resource management in OFDMA femtocell networks

Mohammad ZAREI[1,*], **Behrouz SHAHGHOLI GHAHFAROKHI**[2], **Mehdi MAHDAVI**[1]
[1]Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan, Iran
[2]Department of Information Technology Engineering, University of Isfahan, Isfahan, Iran

Received: 11.10.2015 • **Accepted/Published Online:** 21.07.2016 • **Final Version:** 29.05.2017

Abstract: Recent evolutions in mobile networks have led to increased resource demands, especially from indoor users. Although recent technologies such as LTE have an important role in providing higher capacity, indoor users are not satisfied adequately. Femtocell networks are one of the proposed solutions that support high data rates as well as better indoor coverage without imposing heavy costs to network providers. However, interference management is a challenging issue in femtocell networks, mainly due to dense and random deployment of femto access points (FAPs). Therefore, distinct radio resource management (RRM) methods are employed to ensure acceptable levels of call dropping/blocking probability and spectral efficiency. However, the mobility of mobile users is an important issue in resource management of femtocell networks that has not been considered adequately. In this paper, we propose an algorithm that predicts the resource requirements of FAPs regarding mobility of their users and allocates the resources to the FAPs based on an extended load-based RRM algorithm that prioritizes handoff calls to incoming calls. Simulation results illustrate that the proposed method has shown lower call dropping probability and higher spectral efficiency compared to the benchmark algorithms.

Key words: Femtocell networks, resource allocation, mobility prediction

1. Introduction

Femto access points (FAPs) are low-power and low-cost base stations in heterogeneous cellular networks that provide higher coverage and quality of service (QoS) for indoor user equipment (UE) [1]. Radio resource management (RRM) is an important issue in heterogeneous networks. Given that the FAPs share the same resources with the macro base station (MBS) and also the other FAPs, RRM should mitigate the interference level more carefully [2]. FAPs can be employed in different access modes, namely open, closed, and hybrid access. In this paper, we assume an open access mode where all cellular users are allowed to use the FAP.

Several studies investigated the RRM problem in femtocell networks. The scheme named FERMI [3] uses measurement-driven triggers to separate users that require just link adaptation from those that require resource isolation, in a WiMAX network. The authors proposed a mechanism for joint scheduling of both types of users in the same time frame. Afterwards, an efficient algorithm was employed to determine fair resource allocation based on graph theory regarding utilization. The adaptive clustering heuristic algorithm (ACHA) [4] uses clustering of femtocells to reduce co-tier interference by proper subchannel and power allocation. It is

*Correspondence: m.zarei@ec.iut.ac.ir

noteworthy that in all of the above works, users' mobility was ignored while it may lead to mitigation of users' QoS and reduction of resource utilization.

Mobility of users is one of the important parameters in femtocell networks that affects the users' QoS and resource utilization, especially in open access FAPs. Since the coverage of the FAPs is limited, UEs can likely move out of the coverage of a FAP in a short time. Therefore, the number of handovers is higher compared to macrocell networks [5]. The authors of [6] proposed an admission control method that avoids handoff overheads. Accordingly, supposing the femtocell extended area around the typical femtocell area, fast users in this area are connected to the MBS whereas slow users are associated with the FAP or the MBS that was already connected to it. However, this method does not consider the load changes arising from mobility in resource allocation of FAPs. The authors of [7] proposed joint resource allocation and power control considering the users locations and demands. Their method maximizes the network throughput and minimizes the interference level by means of linear programming.

Mobility prediction is an efficient technique that assists RRM. When the handoff procedure is initiated, if the new FAP does not have enough resources to support the handoff call, the call will be dropped. However, mobility prediction could be exploited to reserve the radio resources in target access networks towards reducing the call dropping probability and improving resource utilization. There are recent studies such as [8–12] that present predictive methods for femtocell networks. To improve the physical cell identity (PCI) collision problem, [8] introduced a dynamic PCI allocation algorithm based on Markov chains to anticipate the high handover requested FAPs and assign a specific number of PCIs to these FAPs while the other FAPs share the remaining PCIs. The approach in [9] presents offiine association control algorithms for femtocell networks to improve the association control problem using mobility prediction. Similarly, the authors of [10] proposed a predictive approach to reduce the handoff delay. In their algorithm, the reference signal received power is anticipated using time series analysis to activate layer 3 handoff prior to the layer 2 handoff procedure. In [11], the authors introduced an algorithm that determines temporary FAP UEs by predicting the next locations of mobile UEs to reduce the handoff overhead. On the other hand, the proposed solution in [12] presents an adaptive recursive least square algorithm in order to predict the future received signal strength (RSS) samples of the target FAPs and the current serving FAP. Thereafter, the list of candidate FAPs is attained based on the estimated signal to interference plus noise ratio (SINR) and predicted RSS, and the FAP that results in the highest throughput is selected.

Regarding the above approaches, mobility prediction has been exploited in order to improve handoff performance. However, these methods have not benefited from mobility prediction to improve the resource reservation and interference management among FAPs. The authors in [13] studied a new approach that improves variable bit rate video traffic in downlink transmission. This scheme exploits a resource reservation algorithm and a handover utility function that take the future video users' connections into account. Similarly, a resource reservation and call admission control mechanism was designed in [14] to specify whether a newly arriving call or a handoff should be accepted and to preserve the resources for probable handoff connections before those take place. In this respect, the disconnection probability of the handoffs is derived based on instantaneous mobility characteristics such as location and speed. However, it must be noted that the method tries to preserve the quality of a connection during mobility of the user without regarding the performance and resource utilization of the overall network, considering long-term history and mobility patterns.

To this end, in this paper we propose a resource allocation mechanism for orthogonal frequency division multiple access (OFDMA) femtocell networks in open access mode that predicts the future load of the FAPs

and addresses a resource allocation algorithm to preserve the radio resources of all the FAPs regarding predicted loads. In contrast to the existing algorithms, such a goal is achieved by utilizing mobility prediction based on users' movement histories, so the FAPs have enough resources for probable handoff calls.

The reminder of this paper is organized as follows: Section 2 introduces baseline methods. Section 3 presents system model and assumptions. Section 4 explains our mobility-aware algorithm and Section 5 presents simulation results. The paper is finally concluded in Section 6.

2. Background

In this section, two algorithms that are employed in the proposed method are introduced. Moreover, the ACHA method that is used as the benchmark method is described in more detail.

2.1. FERMI method

FERMI [3] is a resource management algorithm that includes the following parts:

- Client categorization: The FERMI algorithm uses measurement-driven triggers to classify users that require only link adaptation (LA) and reuse of the frequency resources (class LA) from those that require resource isolation (class ISO). Users of class LA have weak interference from other FAPs. If the SINR level of LA users is above a predefined threshold, changing to a lower level modulation and coding scheme (MCS) could be adequate to mitigate the interference. However, users of class ISO have strong interference. As a result, decreasing the MCS level is not adequate and resource isolation is considered for alleviating the interference.

- Frequency domain isolation: According to the preceding step, for users that are in ISO, the resources are isolated in each femtocell in the frequency domain. The power transmitted by a FAP is divided over its OFDM subchannels. With a lower number of subchannels, the average power per subchannel increases and so the FAP is permitted to apply a higher level MCS for interference compensation. If more FAPs have interference in a domain, the subset of available subchannels per FAP is decreased, which leads to an increase of the average power. Thus, the throughput per subchannel and subsequently the network capacity increases.

- Zoning: The algorithm presents a frame structure for both types of users. The zoning method determines the frame ratio (in symbols) for both types of users in each FAP.

- Resource allocation: The FERMI method proposes a graph-based resource allocation method. It applies novel algorithms to allocate subchannels to interfering FAPs considering a weighted max-min fairness model. The FERMI resource allocation method is not mobility-aware and only allocates the resources based on current load of FAPs. As will be seen, an extension of the FERMI resource allocation method will be proposed in this paper. In the proposed method, the resources are allocated to the FAPs based on predicted loads and so the number of subchannels allocated to the FAPs is closer to the number of their required subchannels in the near future.

2.2. Q-FCRA algorithm

QoS-based femtocell resource allocation (Q-FCRA) [15] is a cluster-based resource allocation algorithm that takes user prioritization into account. In this paper, we employ the Q-FCRA clustering method in order to

eliminate centralized resource management and improve scalability by devolving resource allocation task to cluster heads (CHs). To indicate the CHs, each FAP sends its interference degree (number of interfering FAPs) to its one-hop neighbors at first. The FAP with the highest interference degree among its neighbors is the CH. Otherwise, the neighboring FAP that has the highest interference among its one-hop neighbors is adopted as the CH. If the FAP has no CH between its one-hop neighbors, i.e. all of the neighboring FAPs are members of other CHs, this FAP also becomes a member of the neighboring cluster that has the CH with maximum interference degree. Hence, the Q-FCRA clustering algorithm tries to classify FAPs in clusters such that more overlapping FAPs are in the same cluster. The CH is one of the FAPs in a cluster that has maximum aggregated interference with other FAPs that have formed the cluster.

2.3. ACHA algorithm

The ACHA algorithm [4] classifies FAPs in distinct clusters such that those that have no mutual interference are placed in the same cluster and so reuse the same subchannels. The number of subchannels assigned to the FAPs of a cluster is determined with respect to the number of FAPs in that cluster. Like the FERMI method, this scheme does not consider users' mobility and only allocates the resources based on current load of FAPs. Moreover, this method allocates the resources to the FAPs equally and does not attend to the resource demands of the FAPs. This assumption results in reduction of users' QoS and resource utilization.

3. System model and assumptions

The proposed algorithm is concentrated on downlink resource allocation for OFDMA-based femtocell networks.

The assumed network consists of a macrocell and M FAPs, $F = \{f_1,\ f_2 \dots f_M\}$. We employ an orthogonal channel assignment that eliminates the cross-layer interference between femtocells and the macrocell. Regarding the total number of subchannels as N, the number of subchannels dedicated for the FAPs and the macrocell are N_f and N_m, respectively. The following relationship is thus established:

$$N = N_f + N_m \tag{1}$$

The transmission power of each FAP is assumed to be P_T.

We use the set $U = \{u_1,\ u_2 \dots u_K\}$ to denote the UEs in the network where K is the number of UEs. Each user is only connected with one FAP or MBS. As shown in Eq. (2), C_{mh} indicates the attachment of the hth UE (u_h) to the mth FAP (f_m).

$$C_{mh} = \begin{cases} 1, & u_h \text{is attached to } f_m \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

The allocation matrix A_{mn} illustrates the allocation of subchannel n to f_m as shown below:

$$A_{mn} = \begin{cases} 1, & \text{if subchannel } n \text{ assigned to } f_m \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Denoting the channel gain between f_j and the UEs associated with f_i as $L_{i,j}$, the SINR value of the user of f_i assuming that it only receives interference from FAP j is given by $SINR_{i,j}$, which is calculated as:

$$SINR_{i,j} = \frac{L_{i,i} \times P_T}{L_{i,j} \times P_T + \sigma^2} \tag{4}$$

In Eq. (4), σ^2 is the power of the additive white Gaussian noise.

Based on the given femtocell network, the interference graph $G = (V, E)$ is constructed. The vertex set is represented by $V = \{v_1 \ldots \ v_M\}$ where each vertex denotes a FAP. $E$is the bidirectional edge set. The interference relationships between the FAPs and the UEs can be simplified to interference relationships between the FAPs. Hence, the nodes v_i and $v_j \in V$ $(i \neq j$, and $i, j \in \{1, 2 \ldots M\})$ are connected by an edge if and only if the minimum of $SINR_{i,j}$ and $SINR_{j,i}$ is lower than an SINR threshold, Γ, as shown in Eq. (5).

$$E_{ij} = E_{ji} = \begin{cases} 1, & \text{if } \min(SINR_{i,j}, SINR_{j,i}) < \Gamma \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

3.1. Proposed method

In this section, the proposed method is precisely described. The purpose of this method is providing a dynamic load-based RRM for femtocell networks and it considers the mobility of users to predict the load and allocate the resources based on prediction results. According to Figure 1, the proposed method consists of two modules: the location prediction module, which is located in UEs, and the resource allocation module, which is located in CH FAPs. We assume that each FAP executes the Q-FCRA clustering algorithm and so the clusters were formed earlier.

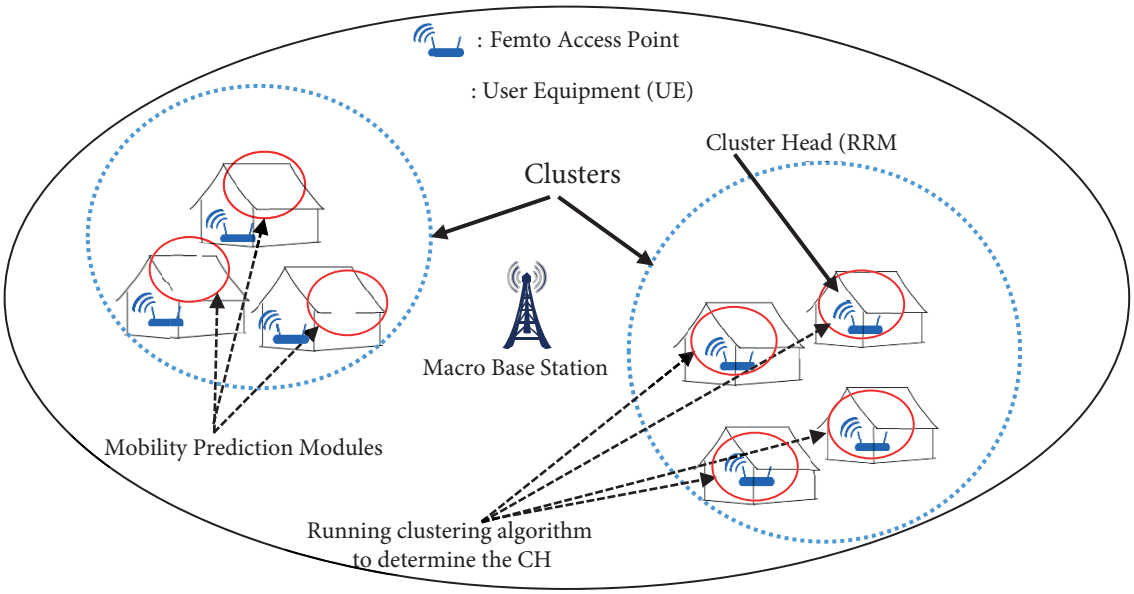

Figure 1. The clustering, location prediction, and RRM modules in proposed method.

3.2. Location prediction module

In this section, we discuss the location prediction module, which is located in UEs. Therefore, each UE periodically predicts its next place and reports it to the resource allocation module. In order to apply the location prediction algorithm, we model the network environment as a grid where each UE is located in one of the grid locations regarding GPS positioning information. The grid is shown by $LOC = \{loc_1, \ldots, loc_{N_g}$, where N_g is the number of grid locations. According to Figure 2, we model the mobility of a user as a sequence of grid locations. In the location prediction module, we exploit the profile-based prediction algorithm that was proposed in [16]. This algorithm tries to predict the next location(s) to be probably visited by the mobile

user, using the list of local profiles. However, as the method of [16] does not always respond, first/second-order Markov-based predictors are also exploited by the module (inspired by [17]).

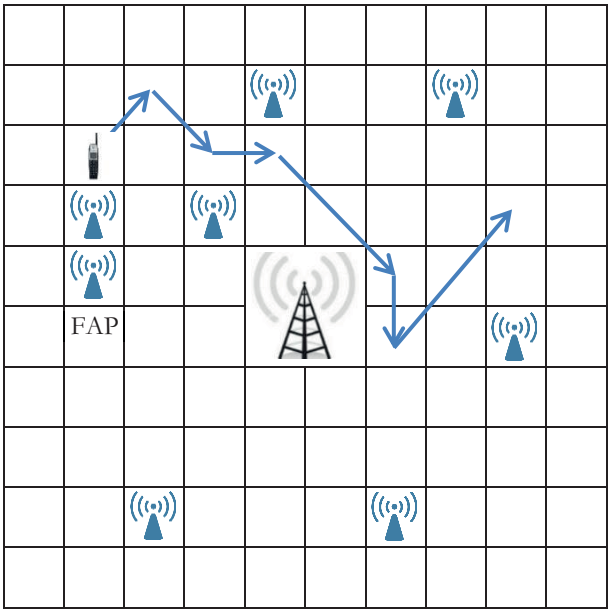

Figure 2. The mobility of users considered as a sequence of grid locations.

For location prediction, each UE is responsible for preparing the mobility profile of its user. A profile is determined as a set of possibly similar sequences of grid locations (from a source location to a destination one) traveled to by the mobile user.

The main prediction mechanism is based on identifying the next local profile $f_i^u \in F^u$, $1 \le i \le NP^u$, to be possibly followed by the mobile user, based on its current location. For that, let $loc_c \in LOC$ and $F_c^u \in F^u$ be the current location and the set of all local profiles of the mobile user containing loc_c, respectively. If $loc_{c-1} \in LOC$ is the previous location visited by the user, the probability P_f that the mobile user follows the local profile, $f_i^u \in F_c^u$, is as follows [16]:

$$P_f = P\left(f_i^u \mid loc_{c-1}\right) = \frac{S_{f_i|loc_{c-1}}}{\sum_{j=1}^{NP_c^u} S_{f_j|loc_{c-1}}} \tag{6}$$

In Eq. (6), NP_c^u is the number of profiles containing the current location loc_c, and $S_{f_k}|loc_{c-1}$ is the number of times that profile f_k^u has been followed by the mobile user u when the previous visited location belongs to f_i^u, i.e. $loc_{c-1} \in f_i^u$. Then, in every profile, the next locations are regarded and their corresponding probabilities are considered to anticipate loc_{c+1} [16].

In the mentioned procedure, if the profile list is empty, the model will not be able to perform predictions. An empty list means that no profile has been pursued by the mobile user that contains the current location, loc_c. In this case, [16] proposed to exploit Markov-based predictors such as the one in [17]. Each UE obtains the probabilities of being in the next likely locations and reports them to its FAP. The above procedure reruns whenever the UE changes its location in the assumed grid.

4. Resource allocation module

In this section, the probabilistic load-based resource allocation scheme is presented. Centralized RRM approaches are not scalable and suffer from low performance whenever the number of FAPs increases. Therefore, the clustering method of [15] that was mentioned in Section 2.2 is used to cluster the FAPs and then the proposed resource allocation mechanism is performed in each of the CHs. The pseudocode of Table 1 describes the proposed algorithm in detail. Moreover, Figure 3 and Table 2 show the proposed resource allocation method by an example assuming a particular case that only considers one cluster.

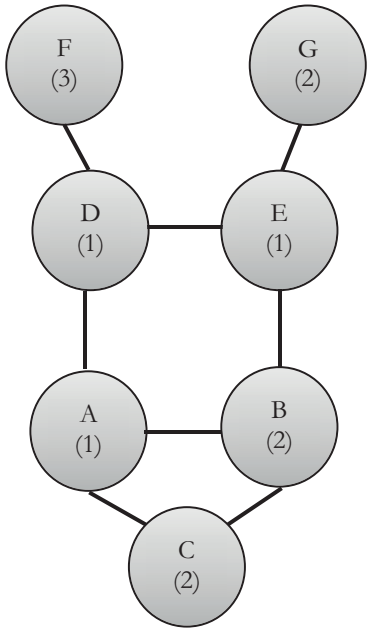

Figure 3. An example of a constructed graph with triangulation edge and first selected maximal clique. The loads are shown in parentheses and it is assumed that N_f is 20.

The algorithm is an extension of the FERMI resource allocation algorithm [3], a demand-based resource allocation mechanism that aims at assigning appropriate numbers of subchannels to the FAPs regarding their demands in such a way that resource utilization is maximized. However, in contrast to the baseline algorithm, the effective demands of FAPs and also the MBS are determined based on mobility prediction results. Accordingly, for any UE, u_k in each epoch (T), the next possible locations, and their arrival probabilities are gathered (output of location prediction module, which is reported to the FAP). The effective demand of each grid location, loc_n, is determined using Eq. (7).

$$D_{eff}(n) = \sum_{k=1}^{K} P_{k,n} \times d_k \tag{7}$$

In Eq. (7), $P_{k,n}$ and d_k are the probability of arrival of u_k into the grid location, loc_n, and the resource demand of u_k, respectively.

Then, for each FAP i in cluster r (where the algorithm is executing in its cluster head), the load $l_{i,r}$ is calculated by collecting the effective demands of the underlying grid locations assuming a coverage distance (d_{th}) for FAPs (lines 5–13). The probable loads of the grid locations that are not covered by FAPs are considered as the MBS load, L_{mbs} (line 17).

Table 1. Proposed mobility and load aware resource management algorithm.

Input: $G_r^+ = (V_r, E_r^+), \forall cl_r \in CL;\ d_k, \forall u_k \in U;\ d_{th}$
Output: Number of allocated subchannels to the base stations

1: Repeat in every epoch time, T:
2: Run location prediction module that it returns $P_{k,n}, \forall u_k \in U, \forall loc_n \in LOC$
3: Determine load of each location index denoted by $loc_n \in LOC$:
$D_{eff}(n) = \sum_{k=1}^{K} P_{k,n} \times d_k$;
4: for each cluster $cl_r \in CL$
5: Set value of $l_{i,r} \in l_r$ to zero, $\forall i: v_{i,r} \in V_r$;
//Associate load of location indices to vertices (FAPs):
6: for all $loc_n \in LOC$
7: find nearest FAP to loc_n, i.e. $v_{p,r}$;
8: if distance $(v_{p,r}, loc_n) \leq d_{th}$
9: $l_{p,r} = l_{p,r} + D_{eff}(n)$;
10: $D_{eff} \leftarrow D_{eff} \backslash D_{eff}(n)$;
11: $LOC \leftarrow LOC \backslash loc_n$;
12: endif
13: endfor
14: Determine all the maximal cliques in G_r^+, as denoted by $C_r = \{c_{1,r}, \ldots, c_{m_r,r}\}$;
15: Determine load of each maximal clique in G_r^+:
$L_{j,r} = \sum_{i:v_{i,r} \in c_{j,r}} l_{i,r}, \forall c_{j,r} \in C_r$;
16: endfor
17: Associate load of remaining location indices to MBS
$L_{mbs} = \sum_{n:loc_n \in LOC} \sum_{k=1}^{K} P_{k,n} \times d(k)$;
18: Specify the proportion of resources for FAPs and MBS:
$N_m = \left\lfloor \frac{L_{mbs}}{L_{mbs} + \max_{j:\, c_{j,r} \in C_r, \forall cl_r \in CL} L_{j,r}} . N \right\rfloor$;
19: $N_f = N - N_m$;
// Allocate subchannels to FAPs and MBS:
20: for each cluster $cl_r \in CL$
21: Set $UV_r = V_r$ and $AV_r = \emptyset$ where UV_r is unallocated vertices and AV_r is allocated vertices in cluster r;
22: $\forall c_{j,r} \in C_r, R_j = N_f$;
23: Determine pairs : $< s_i = max_{j:v_{i,r} \in C_{j,r}}\{L_{j,r}\}, t_i = \sum_{i:v_{i,r} \in C_{j,r}} 1, \forall v_{i,r} >$;
24: Determine initial allocation:
$AF_{i,r} = min_{j:v_{i,r} \in c_{j,r}} \left\lfloor \frac{l_{i,r} \times R_j}{L_{j,r} = \sum_{k:v_{k,r} \in c_{j,r}} l_{k,r}} + 0.5 \right\rfloor, \forall v_{i,r} \in UV_r$
25: while $UV_r \neq \emptyset$ do
26: Pick unallocated vertex with maximum $(s_i + t_i)$ value:
$v_{o,r} = argmax_{i:v_{i,r} \in UV_r}(s_i + t_i)$;
27: Allocate $AF_{o,r}$ subchannels to $v_{o,r}$;
28: $UV_r \leftarrow UV_r \backslash v_{o,r}$;
29: $AV_r \leftarrow AV_r \cup v_{o,r}$;

Table 1. Continued.

30: Update remaining resource: $R_j = R_j - AF_{o,r}, \forall j: v_{o,r} \in c_{j,r}$;
31: Remove $v_{o,r}$from cliques: $c_{j,r} \leftarrow c_{j,r} \backslash v_{o,r}, \forall j: v_{o,r} \in c_{j,r}$;
32: Update $L_{j,r}, \forall L_{j,r} \in L_r$ and $(s_i, t_i), \forall v_{i,r} \in UV_r$;
33: Update allocation $AF_{i,r}$ $\forall v_{i,r} \in UV_r$
34: end while
35: endfor
//Allocate specified resources to users regarding priority of handoff calls:
36: Repeat following steps once for handoff calls and another time for incoming calls
37: for all $u_k \in U$
38: if u_k is connected to FAP, $f_{i,r}$ $(B_{kir} = 1)$
39: if $AF_{i,r} \geq d_k$
40: allocate d_k subchannel to u_k;
41: $AF_{i,r} = AF_{i,r} - d_k$;
42: endif
43: elseif u_k is connected to MBS
44: if $N_m \geq d_k$
45: allocate d_k subchannel to u_k;
46: $N_m = N_m - d_k$;
47: endif
48: endif
49: endif
50: endfor

Table 2. The number of assigned subchannels for vertices (FAPs) of A to F.

Vertices	Number of assigned subchannels
A	Min(4,5,7) = 4
B	Min(8,11) = 8
C	Min(8,5) = 5
D	Min(8,11,7) = 7
E	15
F	8
G	13

In the mentioned example, the supposed probable future loads are shown in nodes of the interference graph of Figure 3. It is assumed that the probable loads of FAPs have been calculated based on previously given location prediction results.

By accomplishing a triangulation process (adding edges to triangulate a graph as shown by dotted lines in Figure 3), the new graph $G_r^+ = (V_r E_r^+)$ is generated from the interference graph of FAPs, G. Afterwards, all maximal cliques, $c_{j,r}$, are determined (for each $j \in \{1..m_r\}$ where m_r indicates the number of maximal cliques in cluster r). Then the method of [18] is used to determine the aggregate load of each maximal clique, i.e. $L_{j,r}$, by adding the loads of all vertices in that maximal clique (lines 14 and 15 of the proposed algorithm).

Now the fraction of resources used by MBS and FAPs (N_f and N_m) must be determined. The FERMI algorithm determines these values proportional to the load of MBS and the maximum load of maximal cliques (lines 18 and 19). N_f and N_m are calculated from the following equations:

$$N_m = \left\lfloor \frac{L_{mbs}}{L_{mbs} + L_{j,r}} \times N \right\rfloor \tag{8}$$

$$N_f = N - N_m \tag{9}$$

Then, in line 21, we place every vertex of cluster cl_r in V_r.

Focusing on our example, we have four maximal cliques (where one of the maximal cliques has been indicated by a circle in Figure 3 and the load of this maximal clique is 5). In Figure 3, we assume that N_f is equal to 20, which is equal to the number of available subchannels for each maximal clique.

Afterwards, the pairs (s_i, t_i) are determined for each vertex (line 23) and the vertex (FAP) with the highest value of $s_i + t_i$ is selected (line 26). Here, s_i refers to the highest load of the different maximal cliques that this vertex (i) belongs to and t_i is the number of cliques that this vertex belongs to. Accordingly, in Figure 3, vertex A is selected (since $s_A = 5 \, and \, t_A = 3$).

A weighted max-min fair allocation is then adopted from FERMI [3] to indicate the number of subchannels for the above selected vertex. The number of subchannels assigned to vertex i ($AF_{i,r}$) is obtained from the minimum value of the ratio of $l_{i,r}$ (load of the FAP) to $L_{j,r}$ (load of the jth maximal clique) between all maximal cliques that i belongs to as shown in Eq. (10), where R_j is initially set to N_f.

$$AF_{i,r} = min_{j:\, v_{i,r} \in c_{j,r}} \left\lfloor \frac{l_{i,r} \times R_j}{L_{j,r}} + 0.5 \right\rfloor \tag{10}$$

As an example in Figure 3, vertex A is in 3 maximal cliques and the loads of these maximal cliques are 5, 4, and 3, so the weighted max-min fair value for vertex A is calculated as below, recalling that R_i is 20:

$$AF_A = \min \left(\left\lfloor \frac{1 \times 20}{5} + 0.5 \right\rfloor \left\lfloor \frac{1 \times 20}{4} + 0.5 \right\rfloor \left\lfloor \frac{1 \times 20}{3} + 0.5 \right\rfloor \right) = 4. \tag{11}$$

Therefore, 4 subchannels are allocated to FAP A at this step.

After that, the number of allocated subchannels is decreased from the number of available resources of the maximal clique, i.e. R_j. The corresponding vertex is also emitted from its maximal clique. Finally, the load of maximal cliques, pairs (s_i, t_i), and $AF_{i,r}$ get updated for all remaining vertices (e.g., vertices B, C, D, E, F, and G in Figure 3). This procedure is performed for each vertex of the cluster (line 20) subsequently. Consequently, the number of subchannels assigned to the FAPs of Figure 3 is finally determined as shown in Table 2.

It is noteworthy that as the proposed method gives higher priority to handoff calls, the subchannels that are assigned to the base stations are first allocated to the handoff calls rather than new incoming calls (lines 36–50).

5. Performance evaluation

In this section, we evaluate the efficiency of our proposed method through a simulated network. The system parameters and simulation assumptions are presented in Table 3. In each simulation scenario, the FAPs are

located in the center of 10 m × 10 m apartments that are considered as grid locations. The number of FAPs is considered between 10 and 70 (when the number of UEs is 50) and the number of UEs is considered between 10 and 70 depending on each scenario. The path loss is calculated using Eq. (12) at distance d meters from the transmitter [19]. Also, each UE requests a random number of subchannels between 1 and 5 according to a uniform distribution.

$$\text{Pathloss} = 37 + 30 \times \log_{10}(d) \tag{12}$$

We use the NCSU human mobility trace [20], which is collected from various sites. Among them, we have used a university campus (KAIST) mobility model as the training and test data to evaluate the performance of our approach.

Table 3. Simulation parameters.

Values	Parameters
13 dBm	FAP power
43 dBm	MBS power
5 MHz	Bandwidth
37 + 30 × $\log_{10}$(d)	Path loss
9 dB	Receiver noise figure
25	Number of subchannels
[10, 60]	Number of FAPs
[10, 70]	Number of users
Grid model	FAP layout
200 m × 200 m	Network area
10 m × 10 m	Grid size
Center of the area	MBS location
Center of the grid locations	FAP location
Uniform distribution in [1,5]	Number of subchannels demanded

It should be noted that according to [21], the call duration time has a log-normal distribution with a mean of 1 min. The simulation results are obtained for 200 min and due to random places of FAPs, 30 random scenarios with variable seeds are tested and the average of the simulation results is reported. In every scenario, the possibility of deploying a FAP in each grid location (apartment) is determined according to a uniform distribution.

5.1. Simulation results

The simulation results compare the demand-based resource allocation algorithm of FERMI [3] and the ACHA method [4] to the proposed mobility-aware method.

In the location prediction module, for each user, a list of locations that the UE is likely to be at in the future (and their presence probabilities) is generated at every time step. Given that we know the user's next location from the dataset, Figure 4 demonstrates the prediction accuracy versus the time steps. According to Figure 4, by increasing the grid resolution, the number of location indices within the network increases, which results in more precision in locating users. In the remaining evaluations, the resolution of 10 m × 10 m apartments is considered.

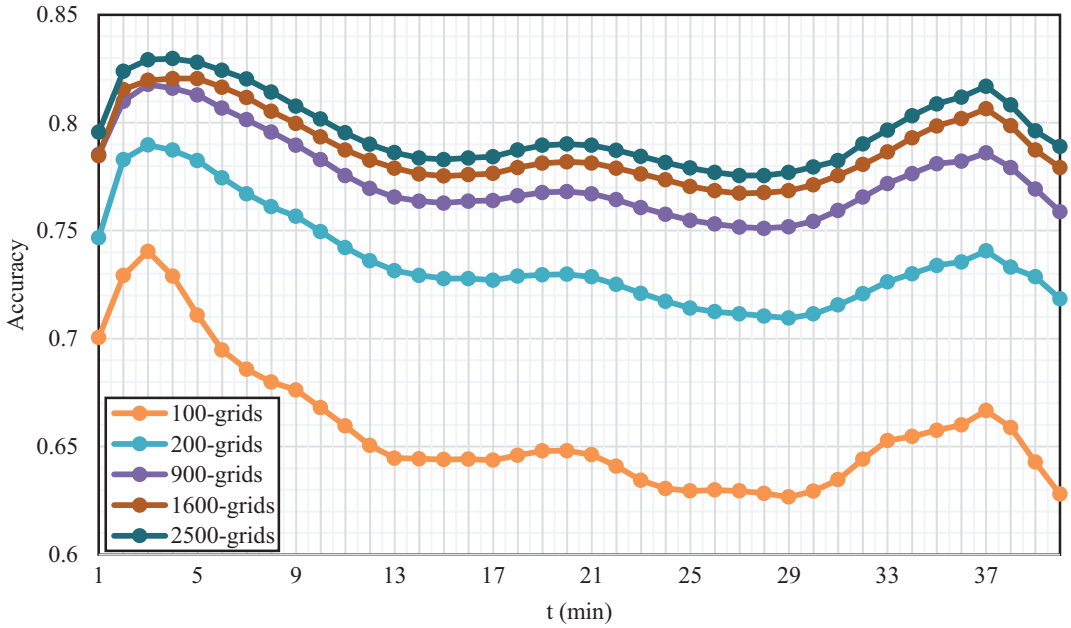

Figure 4. Mobility prediction accuracy.

Network capacity and subchannel utilization are plotted versus the number of FAPs and the number of users for all mentioned methods as shown in Figures 5 and 6, respectively. Eqs. (13) and (14) show the definition of network capacity [22] and subchannel utilization metrics, where BW_{sc} is the bandwidth of a subchannel (here, 180 kHz) and A_{ik} illustrates the allocation of subchannel i to FAP f_k.

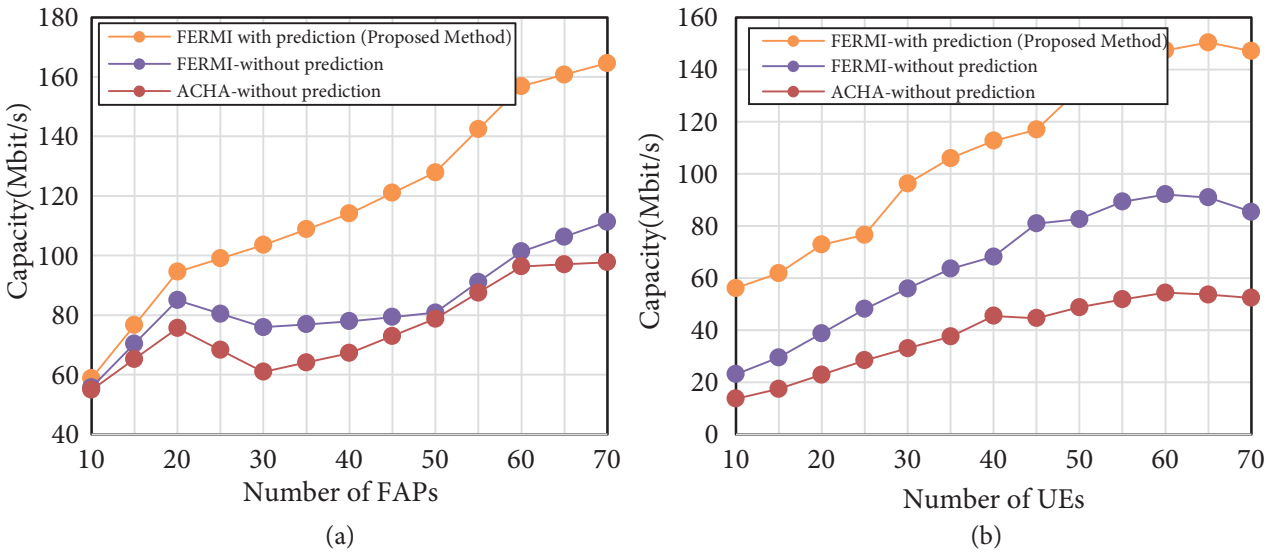

Figure 5. Network capacity comparison versus the number of FAPs (a) and the number of users (b).

$$\text{Capacity} = \sum_{k=1}^{K}\sum_{i=1}^{N} A_{i,k} BW_{sc} \log_2(1 + SINR_{i,k}) \tag{13}$$

$$\text{Utilization} = \frac{\text{The number of subchannels that are used by UEs}}{\text{Total number of subchannels}} \tag{14}$$

As can be seen in Figures 5a and 5b and 6a and 6b, the network capacity and utilization of the proposed algorithm are higher compared to the benchmark approaches. The reason is that, in the proposed method, the resources are allocated to the FAPs based on the predicted status of the future load and therefore the number of subchannels that are assigned to the FAPs is closer to the amount of their required channels in the near future. However, in traditional methods that only consider the static load, the resources may be underutilized due to the mobility of the load over time. Also, as ACHA is not a load-based algorithm, it cannot discriminate among different loads and so inappropriate allocation of radio resources reduces the capacity and utilization more.

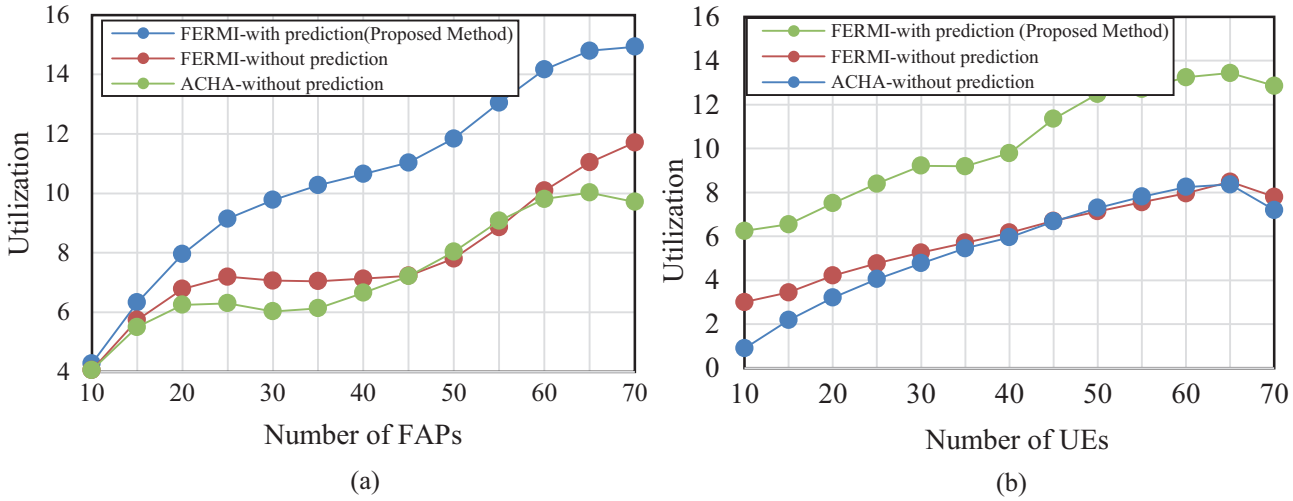

Figure 6. Resource utilization comparison versus the number of FAPs (a) and the number of users (b).

As can be observed in Figures 5a and 6a, when the number of FAPs is low, users are less likely to be connected to the FAPs. As a result, the resource reuse and thus the utilization is low. Also, as users have fewer resources overall, network capacity is low, too. Increasing the number of FAPs, more UEs can connect to the FAPs. Thus, resource reuse increases and UEs connect through nearby base stations with better channel conditions, which will increase network capacity and utilization. Similarly, in Figures 5b and 6b, it is shown that by increasing the number of UEs, the network capacity and resource utilization increase until they reach steady points. This is due to the fact that increasing the number of users causes higher exploitation of network resources, but finally due to the saturation of network resources, further increase in UEs does not increase the utilization as further requests are blocked due to the lack of radio resources.

The call dropping probability (CDP) and call blocking probability (CBP) are other important metrics that are used for evaluation of the proposed method. Call dropping and call blocking occur whenever a base station has no free subchannel to allocate to a mobile user. Here, call blocking refers to blocking new incoming calls due to the lack of available subchannels, and call dropping refers to the drop of ongoing calls due to the lack of subchannels in target cells after handover of the UEs. The goal of almost all resource management methods is to lower the CDP and CBP while maintaining higher bandwidth utilization. Figures 7 and 8 represent CDP and CBP versus the number of FAPs and the number of UEs where CDP and CBP are calculated from Eqs. (15) and (16), respectively.

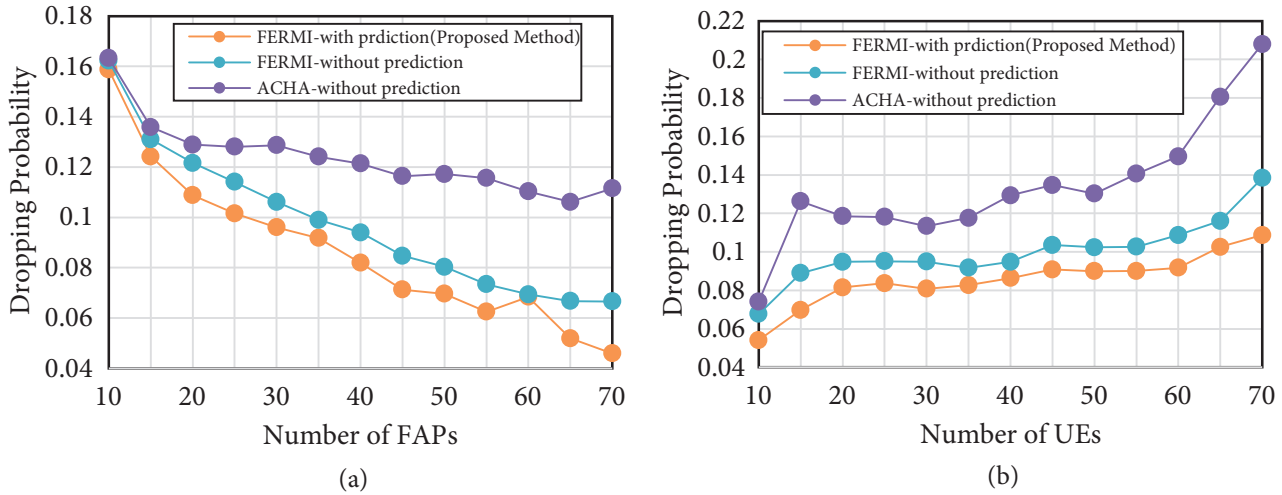

Figure 7. Call dropping probability comparison versus the number of FAPs (a) and the number of users (b).

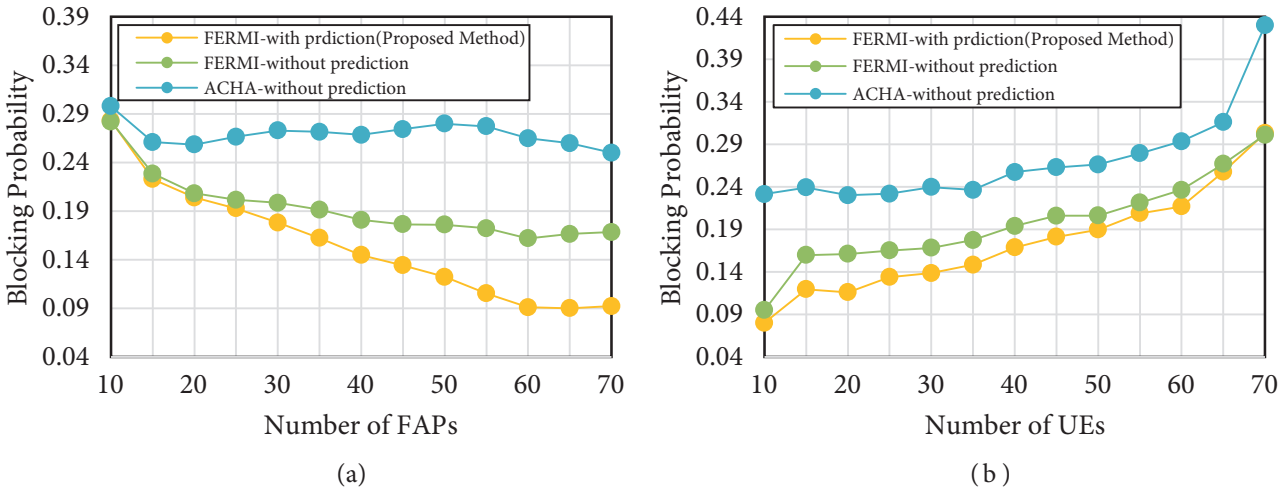

Figure 8. Call blocking probability comparison versus the number of FAPs (a) and the number of users (b).

$$CDP = \frac{\text{Number of dropped calls}}{\text{Total number of handoff calls}} \tag{15}$$

$$CDP = \frac{\text{Number of blocked calls}}{\text{Total number of incoming calls}} \tag{16}$$

With respect to Figures 7a and 7b and 8a and 8b, we see that the proposed algorithm has lower CDP and CBP compared to the FERMI resource allocation method. This is due to the fact that the location prediction module predicts the resource requirements of FAPs. Hence, the target FAPs have possibly adequate resources for handoff calls, which reduces the CDP. Consequently, there is more chance of even accepting new calls with the remaining channels, which decreases CBP, too. ACHA does not consider users' demands and so the CDP and CBP of ACHA are higher than those of the other methods.

As shown in Figures 7a and 8a, by increasing the number of FAPs, the CDP and CBP decrease. The

reason is that by increasing the number of FAPs, the amount of resources that are reused within the network increases. In contrast, in Figures 7b and 8b, it is shown that by increasing the number of users, the CDP and CBP are increased. According to the figures, when the network reaches the saturation state, this increase is more gradual.

It should be noted that when comparing Figures 7a and 7b with Figures 8a and 8b, CDP is slightly lower than CBP as we have prioritized handoff calls over new calls.

5.2. Complexity analysis

The proposed algorithm consists of three parts including FAP clustering, a location prediction module, and a resource allocation module. As the clustering algorithm is selected independently of the proposed method, we do not discuss the complexity of the exploited clustering algorithm. The resource allocation algorithm is executed by each CH in every time step, T. According to [18] the algorithm has a triangulation process such that its time complexity is of O $(|V||E|)$ and a maximal cliques search, which is of O $(|V|)$.

As in [16], implementing the location prediction algorithm at the UEs allows the network to prevent any scalability constraints. This part of complexity is imposed to UEs. The location prediction module uses one of the local user profile-based or Markov-based mechanisms where the complexity of these methods depends on the number of stored locations (due to training of the predictor). In the first scheme, if the total number of stored sequences and path length are equal to N and L, respectively, then the total number of paths is as below:

$$\text{num\_paths} = N - L + 1 \tag{17}$$

As the paths should be mutually compared in order to evaluate the similarity, and regarding the fact that N is much larger than L, the time complexity of training will be O (N^2). Also, for the second method, as any location in the user movement history should be compared to the current location, the complexity is O (N). Thus, depending on the number of user profiles, the time complexity of mobility prediction will be between O (N) and O (N^2). As noted, the prediction is executed every time step Tsimilar to the resource allocation algorithm. Therefore, the value of T must be determined in such a way that not only do the algorithms have enough time, but also the prediction accuracy remains acceptable.

6. Conclusion

In this paper, we proposed a mobility- and load-aware resource allocation algorithm in OFDMA femtocell networks that predicts the resource requirements of FAPs regarding mobility of UEs. Therefore, resources are allocated to the FAPs more efficiently using a load-aware resource management algorithm, which is based on a conventional graph-based method. Through simulation results, we show that our method can achieve significant gain in terms of network capacity, subchannel utilization, CDP, and CBP compared to traditional benchmarks. Furthermore, by prioritizing handoff calls to new calls, we have reduced the CDP.

References

[1] Andrews JG, Claussen H, Dohler M, Rangan S, Reed MC. Femtocells: past, present, and future. IEEE J Sel Area Comm 2012; 30: 497-508.

[2] Liang YS, Chung WH, Ni GK, Chen IY, Zhang H, Kuo SY. Resource allocation with interference avoidance in OFDMA femtocell networks. IEEE T Veh Technol 2012; 61: 2243-2255.

[3] Arslan MY, Yoon J, Sundaresan K, Krishnamurthy SV, Banerjee S. A resource management system for interference mitigation in enterprise OFDMA femtocells. IEEE ACM T Network 2013; 21: 1447-1460.

[4] Li H, Xu X, Hu D, Qu X, Tao X, Zhang P. Graph method based clustering strategy for femtocell interference management and spectrum efficiency improvement. In: IEEE 2010 Wireless Communications Networking and Mobile Computing Conference; 23–25 September 2010; Chengdu, China. New York, NY, USA: IEEE. pp. 1-5.

[5] Xenakis D, Passas N, Merakos L, Verikoukis C. Mobility management for femtocells in LTE-advanced: key aspects and survey of handover decision algorithms. IEEE Commun Surv Tut 2014; 16: 64-91.

[6] Le LB, Hossain E, Niyato D, Kim DI. Mobility-aware admission control with QoS guarantees in OFDMA femtocell networks. In: IEEE 2013 International Communication Conference; 9–13 June 2013; Budapest, Hungary. New York, NY, USA: IEEE. pp. 2217-2222.

[7] Estrada R, Otrok H, Dziong Z, Barada H. Joint BS selection and resource allocation model for OFDMA macro-femtocell networks incorporating mobility. In: IEEE 2013 International Selected Topics in Mobile and Wireless Networking Conference; 19–21 August 2013; Montreal, Canada. pp. 42-47.

[8] Xiao Z, Chen J, Wang D, Li R, Yi K. Interference management via access control and mobility prediction in two-tier heterogeneous networks. J Cent South Univ T 2014; 21: 3169-3177.

[9] Sung NW, Pham NT, Huynh T, Hwang WJ. Predictive association control for frequent handover avoidance in femtocell networks. IEEE Commun Lett 2013; 17: 924-927.

[10] Li H, Ci S, Wang Z. Prediction handover trigger scheme for reducing handover latency in two-tier femtocell networks. In: IEEE 2012 Global Communications Conference; 3–7 December 2012; Anaheim, CA, USA. New York, NY, USA: IEEE. pp. 5130-5135.

[11] Jeong B, Shin S, Jang I, Sung NW, Yoon H. A smart handover decision algorithm using location prediction for hierarchical macro/femto-cell networks. In: IEEE 2011 Vehicular Technology Conference; 5–8 September 2011; San Francisco, CA, USA. New York, NY, USA: IEEE. pp. 1-5.

[12] Yousefi S, Shayesteh MG, Kalbkhani H. Adaptive handover algorithm in heterogeneous femtocellular networks based on received signal strength and signal-to-interference-plus-noise ratio prediction. IET Commun 2014; 8: 3061-3071.

[13] Salhi M, Trabelsi S, Boudriga N. Mobility-assisted and QoS-aware resource allocation for video streaming over LTE femtocell networks. ECTI Trans Electr Eng Electron Commun 2015; 13: 42-53.

[14] Huang CJ, Chen PC, Guan CT, Liao JJ, Lee YW, Wu YC, Chen IF, Hu KW, Chen HX, Chen YJ. A probabilistic mobility prediction based resource management scheme for WiMAX femtocells. In: IEEE 2010 International Measuring Technology and Mechatronics Automation Conference; 13–14 March 2010; Changsha, China. New York, NY, USA: IEEE. pp. 295-300.

[15] Hatoum A, Langar R, Aitsaadi N, Boutaba R, Pujolle G. Cluster-based resource management in OFDMA femtocell networks with QoS guarantees. IEEE T Veh Technol 2014; 63: 2378-2391.

[16] Barth D, Bellahsene S, Kloul L. Mobility prediction using mobile user profiles. In: IEEE 2011 International Modeling, Analysis & Simulation of Computer and Telecommunication Systems Symposium; 25–27 July 2011; Singapore. New York, NY, USA: IEEE. pp. 286-294.

[17] Bellahsene S, Kloul L. A new Markov-based mobility prediction algorithm for mobile networks. In: Aldini A, Bernardo M, Bononi L, Cortellessa V, editors. Computer Performance Engineering. Berlin, Germany: Springer, 2010. pp. 37-50.

[18] Berry A, Blair JRS, Heggernes P, Peyton BW. Maximum cardinality search for computing minimal triangulations of graphs. Algorithmica 2004; 39: 287-298.

[19] ETSI. 3GPP T 36. 92. v. 11. 0. Evolved Universal Terrestrial Radio Access (E-UTRA); FDD Home eNode B (HeNB) Radio Frequency (RF) Requirements Analysis. Sophia Antipolis, France: ETSI, 2012.

[20] Rhee I, Shin, M, Hong S, Lee K, Kim S, Chong S. CRAWDAD Data Set, 2009. Available online at http://crawdad.org/.

[21] Guo J, Liu F, Zhu Z. Estimate the call duration distribution parameters in GSM system based on K-L divergence method. In: IEEE 2007 International Wireless Communications, Networking and Mobile Computing Conference; 21–25 September 2007; Shanghai, China. New York, NY, USA: IEEE. pp. 2988-2991.

[22] Mogensen P, Na W, Kováes IZ, Frederiksen F, Pokhariyal A, Pedersen KI, Kolding T, Hugl K, Kuusela M. LTE capacity compared to the Shannon bound. In: IEEE 2007 Vehicular Technology Conference; 22–25 April 2007; Dublin, Ireland. New York, NY, USA. pp. 1234-1238.