Context-aware and priority-based user association and resource allocation in heterogeneous wireless networks

Mohamad Zalghout, Ayman Khalil, Matthieu Crussière, Samih Abdul-Nabi, Jean-François Hélard

Context-Aware and Priority-Based User Association and Resource Allocation in Heterogeneous Wireless Networks

Mohamad Zalghout[1], Ayman Khalil[1], Matthieu Crussière[2], Samih Abdul-Nabi[1], Jean-Francois Hélard[2]

Abstract

Heterogeneous wireless networks (HWNs) are usually characterized by the integration of cellular networks and wireless local area networks (WLANs) to meet user requirements and enhance system capacity. This paper proposes a user association and downlink resource allocation algorithm in HWNs with users having different priorities. The proposed solution employs contextual information related to the preferences of the users, their requested data rate, and the characteristics of networks. The user preference is translated through a profit function that is based on the received signal quality and the power consumption at mobile terminals (MTs). Accordingly, an optimization problem is formulated to maximize the overall user satisfaction for each priority level. The formulated problem throws firm restrictions to prevent low-priority users from allocating resources utilized by other users with higher priorities. We then propose a novel heuristic approach with polynomial-time complexity to approximate the formulated problem. Furthermore, the system architecture is discussed and a new solution management strategy is proposed to limit the complexity of the algorithm. Simulation results show that the proposed approximation method maintains the nearest performance to the optimal solution.

Keywords: `Power consumption`, `IEEE 802.21 MIH`, `blocking probability`, `binary linear programming optimization`, `user-centric profit`

1. Introduction

With the recent widespread deployment of the fourth generation (4G) wireless communication systems, the fifth generation (5G) mobile and wireless communication technologies are emerging into research fields. It is indicated that the expansion of the wireless data traffic requirements exceeds the capacity growth rate of new wireless access technologies [1]. Since the efficiency of wireless links is approaching its theoretical limits, and the amount of requested data rate is severely increasing, next-generation mobile wireless networks are moving toward heterogeneous architectures usually referred to as heterogeneous wireless networks (HWNs) [1]. In HWNs, users have the right to connect to different types of radio access technologies like long-term evolution (LTE) base stations (BSs) or Wi-Fi access points (APs). Such architecture increases the capacity of the system by reducing the number of users competing for resources at BSs, and supplying those users with better chances to be associated to networks with good channel conditions.

To access the Internet through HWNs, current mobile terminals (MTs) are equipped with multiple wireless access network interfaces. One type of terminals widely used nowadays is that with multiple data interfaces but can benefit from a single interface at a time, usually referred to as a multi-mode terminal. By contrast, multi-homed terminals use multiple interfaces to share the load requested by a single MT. However, a realistic implementation for the multi-homing scenario is still far from deployment and imposes extra complexity on the system. Therefore, the multi-mode terminals are considered. Upon using multi-mode terminals, transferring an ongoing active connection to a new network is probably desired. The transfer in connection could be due to the user mobility, network congestion, user equipment status, *etc.* The process of transferring an active connection between networks is called handover (HO). If the networks that are participating in the HO are of different access technologies, *e.g.* handoff from an LTE BS to Wi-Fi AP, the connection transfer is usually referred to as vertical HO (VHO) [2].

In fact, combining and integrating different types of access technologies in HWNs provides flexible choices for the user to associate with his most preferred available network. In general, users prefer to be associated with the network that provides lower power consumption, better signal quality, better quality of service (QoS), security, *etc.* Consequently, HWNs are usually accompanied with the concept of always best connected (ABC) [3], which is the process of being connected to the best available network at all times. However, the ABC concept is usually taken from the user perspective to rank candidate networks and connect to the best one. Usually, ABC-based network selection algorithms do not consider the limited resources of the networks and the effect of the HO algorithm on the system. Therefore, it is essential in HWNs to pave the way for an

[1]Department of Computer and Communications Engineering, Lebanese International University, Beirut, Lebanon.

[2]Institute of Electronics and Telecommunications of Rennes (IETR) UMR CNRS 6164, INSA Rennes, Rennes, France.

Email address: {mohamad.zalghout, ayman.khalil, samih.abdul-nabi, matthieu.crussiere, jean-francois.helard}@insa-rennes.fr

optimized context-aware ABC scheme that considers both user and network requirements.

1.1. Related Works and Motivations

Extensive research has explored the network selection issue in HWNs. Some studies focus only on one parameter to take HO decisions. In [4] for instance, the HO algorithm selects the network with the highest available bandwidth. In [5], to maximize the MT battery lifetime, the HO algorithm selects the network that requires the lowest power consumption among candidate networks. Other studies consider multiple parameters to rank available networks and select the best one. Usually, each parameter is associated to a weight that indicates its importance among other parameters. The weight of each parameter is set according to the user preferences. Weighted cost function is used in [6], [7], and [8] to rank candidate networks according to the monetary cost, MT power consumption, and QoS-related parameters. However, all previous studies do not consider a system-wide resource allocation and user association solution. Instead, they are designed to satisfy the needs of each user individually.

From a system perspective, several studies with different objectives focus on user association and resource allocation in HWNs. For example, in [9], the proposed solution increases the overall user-centric utility that is based on the per-user throughput. Increasing the per-user throughput has been also used in [10] and [11] as a mean to increase the system throughput. However, the studies [9], [10], and [11] do not consider the amount of data rate requested by each user. In fact, increasing the per-user throughput does not always contribute better satisfaction for users. For example, voice over Internet protocol (VoIP) applications usually request a fixed amount of data rate; increasing the data rate above this amount does not necessarily enhance the performance of the application.

The amount of data rate requested by each user is considered within the optimization functions formulated in studies [12] and [13]. However, the objective is to minimize the amount of time required to satisfy each user traffic demands without considering user-centric welfare. Moreover, the system model and the formulated problems in [12] and [13] do not follow the specific-access-technology resource allocation constraints. Instead, time and frequency resources are assumed to be infinitely divisible. In practice, taking LTE as an example, resources are discrete, and a single resource unit could not be shared between various MTs simultaneously.

The studies [14] and [15] have explored the problem of optimizing the user-centric satisfaction while considering user-demand diversity. However, their proposed optimization function and system model do not follow the specific-access-technology constraints. Instead, a generalized problem formulation is proposed. Furthermore, the optimization problem in [14] do not consider the user preferences.

On the other hand, all previous papers do not take into consideration different user priorities when making decisions. It is common that users in communication systems have different priorities. For example, in mobile networks, users experiencing low long-term transmission rate, or users demanding high QoS, are given higher priority [16]. Moreover, future mobile networks should prioritize the service of emergency applications over the ordinary ones. The authors of [17] have proposed that upon congestion in public safety networks (PSNs) users handoff to LTE system. To ensure reliable service for those users, they are given higher priority among ordinary commercial users. In [18], authors have introduced the concept of degraded utility to deal with different user priorities; additional bandwidth is released to high priority users by degrading the low priority traffic. The authors of [16] have formulated an optimization problem to associate users with different priorities in heterogeneous networks. However, the studies [16] and [18] do not consider specific user-related parameters, and the adopted system model is not realistic. Moreover, both studies do not propose a firm mechanism to prevent low-priority users from allocating resources that could be utilized by users with high priority.

Usually, for the multi-mode MTs, user association is formulated as a binary matching problem. The user association variable is restricted to have a binary value that indicates whether or not a MT is associated to a specific network. However, such problems are known for having an NP-complete complexity which makes the solution intractable. A popular approach used to overcome this issue is to relax the binary association variable into continuous. Then, the solution of the relaxed problem, which usually has a polynomial-time complexity, is used to get the final association decision. In [10], a simple rounding approach is used to convert the fractional association variables into boolean. In [11], the MT connects to the network with the highest fractional association value. However, both solutions are not suitable for the case where MTs request a specific amount of data rate; both approaches could lead to a congested network where the number of available resources is not sufficient to supply each MT with its requested data rate. Moreover, relaxing the binary constraint threatens the optimality of the solution.

Following the goals of the ABC concept that aims at enhancing user satisfaction and considering user preferences, and motivated by the system-wide and priority-based solutions, this paper aims at providing an optimized and priority-based user association and resource allocation scheme to maximize user satisfaction in HWNs. Moreover, this paper explores the heterogeneity of users with different demands and preferences.

1.2. Contributions and Organization

In this paper, we discuss the user association and downlink resource allocation problem in HWNs. MTs in our context have different service priorities, or service levels (SLs), such that MTs with highest priority should experience the best service. In order to be served, each MT should be supplied with its requested data rate, otherwise the MT terminates its ongoing session. Typically, users with high priority should encounter the minimal attainable blockage. In this paper, we optimize the ABC scheme that considers the preferences of users, their priorities, their requested data rates, and the network constraints.

In that perspective, we first formulate a novel binary linear programming (BLP) problem that ensures lower blockage and

better service for high-priority users. The formulated problem exploits different context information that could be user-centric (power consumption, signal quality, and preferences), service-centric (the amount of requested data rate), and network-centric (number of available resources, geographical location, transmission range, *etc.*). The formulated problem throws firm restrictions to prevent low-priority users from allocating resources that could be utilized by other users with higher priorities. Specifically, the algorithm aims at maximizing, for each SL, the user-centric gain which is based on the received signal quality and instantaneous power consumption at the MT. Major contribution in this paper is the description of a novel-heuristic approach to approximate the formulated optimization problem. We compare the performance of our solution to the approach based on relaxing the binary association variable. However, unlike [10] and [11], where both relaxation-based solutions do not account for the data rate requested by MTs, we propose a suitable solution to convert fractional values into binary while supplying each MT with its requested data rate. We also compare the proposed approach to a simple greedy heuristic solution.

Moreover, we discuss the system architecture that is based on the IEEE 802.21 standard. The system is managed by a centralized entity that is responsible for allocating resources and associating users. In addition, a novel solution management strategy is proposed to minimize the number of times the optimization function is processed without affecting the optimality of the solution.

To sum it up, the contributions proposed in this paper could be summarized in the following three main points:

- The formulated optimization problem and the proposed solution management strategy considers users having different priorities.
- The formulated problem also considers user data rate requirements and the network resource allocation constraints.
- A novel solution to approximate the formulated binary optimization problem is proposed and compared to the standard relaxation-based solution.

The rest of this paper is organized as follows. In Section 2, the system model is presented. In Section 3, the optimization problem is formulated, the relaxation-based approach is discussed, and the new approximation-based solution is proposed. In Section 4, the system architecture is discussed and the novel solution management strategy is introduced. Section 5 provides performance evaluation through simulations. Finally, Section 6 concludes this paper.

2. System Model

In this paper, we focus on the downlink resource allocation in a heterogeneous wireless system. The system consists of LTE BSs and Wi-Fi APs with overlapping coverage areas. The set of MTs located within the system is symbolized by $\mathcal{M} = \{1, 2, ..., M\}$. The network sets corresponding to mobile BSs and Wi-Fi APs are denoted by $\mathcal{N}_{\mathcal{BS}} = \{1, 2, ..., G\}$ and $\mathcal{N}_{\mathcal{AP}} = \{G + 1, ..., N\}$ respectively. The total network set is denoted by $\mathcal{N} = \mathcal{N}_{\mathcal{BS}} \cup \mathcal{N}_{\mathcal{AP}} = \{1, 2, ..., N\}$, where $\mathcal{N}_{\mathcal{BS}} \cap \mathcal{N}_{\mathcal{AP}} = \varnothing$. Throughout this paper, "network" n indicates that n belongs to set $\mathcal{N}$, "AP" n is equivalent to $n \in \mathcal{N}_{\mathcal{AP}}$, and "BS" n is equivalent to $n \in \mathcal{N}_{\mathcal{BS}}$. Moreover, MT m in general indicates that m belongs to set $\mathcal{M}$ unless it is needed to be stated otherwise. Note that most of the used variables are defined in Table 1. The available mobile BSs and Wi-Fi APs selected by a given MT are those for which this MT is located in their coverage area. Therefore, we assume that each network n has a circular coverage area with radius R_n, and d_{mn} denotes the distance between the AP or BS n and MT m. The set of all available SLs is denoted by $\mathcal{K} = \{1, 2, ..., K\}$. The data rate in kbps requested by MT m is denoted by Q_m. Since the priority of a MT at a given moment is determined according to the SL, a MT is assigned a single SL k at a given moment. The SL of MT m is denoted by l_m. For simplicity, higher SL indicates higher priority. We define a set θ_k containing all MTs with SL k such that $\theta_k = \{m \in \mathcal{M} : l_m = k\}$.

2.1. Resource Allocation in LTE

We consider the downlink of an LTE BS. The total bandwidth in BS n is divided into C_n sub-channels. Each sub-channel is made up of twelve sub-carriers that are grouped into a resource block (RB) whose total bandwidth is B_n^{RB} kHz. Following similar approach as in [19], the positive channel power gain between MT m and BS n is denoted by H_{mn}. In fact, H_{mn} encompasses the effects of path loss, log-normal shadowing, and antenna gains as large scale fading component (denoted by G_{mn}), and the multi-path Rayleigh fading as small scale fading component (denoted by F_{mn}). In [20], F_{mn} is modeled as an independent exponentially distributed random variable with a unit variance because the envelope of the signal in Rayleigh fading environment is assumed to follow a Rayleigh distribution. Therefore, based on [20], authors in [19] and [21] assume that F_{mn} fluctuates fast enough so that a MT can average it out in its channel measurements. Thus, the long-term signal-to-interference noise ratio (SINR) that is measured by MT m from BS n on a RB is [19]:

$$\overline{SINR}_{mn} = \frac{P_n G_{mn}}{\sum_{i \in \mathcal{N}_{\mathcal{BS}} \setminus n} P_i G_{mi} + B_n^{RB} N_0} \tag{1}$$

where P_n and P_i denote the transmission power on a RB by BSs n and i respectively, N_0 the thermal noise spectral power, and $\mathcal{N}_{\mathcal{BS}} \setminus n$ the set of all BSs except BS n. It is assumed that the allocated power for each sub-channel is predefined. For example, equal power allocation (EPA) could be considered [22]. Therefore, MT m can measure the channel gain for all BSs. Hence, the long-term spectral efficiency in kbps/Hz between MT m and BS n on a RB is [19]:

$$\gamma_{mn} = \log_2(1 + \overline{SINR}_{mn}) \tag{2}$$

where the achievable data rate in kbps on a RB could be calculated by multiplying γ_{mn} by the bandwidth of a RB (B_n^{RB}) and the time duration, then divided by the scheduling interval [19] [23]. Accordingly, we assume that that the transmission time,

or scheduling interval, is divided into T_n^{BS} discrete time fractions, where each RB spanning the interval of one time fraction is identified as a scheduling block (SB). Hence, the total number of SBs at BS n is $U_n = C_n T_n^{BS}$, and u_{mn} denotes the number of SBs that is allocated to MT m if it is connected to BS n. Thus, the long-term achievable data rate (kbps) of MT m in BS n is:

$$r_{mn} = \frac{u_{mn} B_n^{RB} \gamma_{mn}}{T_n^{BS}} \tag{3}$$

While allocating resources, it is more convenient to consider the long-term achievable data rate instead of the instantaneous one, otherwise, the resource allocation algorithm might run upon any degradation in the instantaneous SINR. Moreover, in this paper, the resource allocation algorithm considers the data rate requested by each MT. The data rate requested by MTs could vary dramatically. For example, a MT running a file download application requests data rate much larger than another MT running a VoIP call. Therefore, it is not convenient to allocate a whole RB to MTs requesting low data rate. Hence, the presented system model is adopted. It is beneficial to highlight here that the relative RSS is more inclined towards providing connectivity to the MTs, while the SINR is directly related to the amount of data rate that could be supplied to each MT.

2.2. Resource Allocation in Wi-Fi

In Wi-Fi APs, as in [24], we consider an enhanced version of distributed coordination function (DCF) [25] with a reservation-based medium access control (MAC) protocol. MTs can completely avoid collisions by acknowledging their back-off timer value within the MAC header. Thus, it can be simply seen as if MTs access the AP in a time division multiple access (TDMA) manner. The resource allocation in Wi-Fi APs is also seen as TDMA in [26]. Each MT can occupy the whole bandwidth of AP n, denoted by B_n, in its allocated time slot. The total number of time slots during a scheduling duration is T_n^{AP}, and t_{mn} denotes the number of time slots that is allocated to MT m if it is connected to AP n. Therefore, the data rate (kbps) that is supplied to MT m if it is connected to AP n is:

$$r_{mn} = \frac{r_{mn}^{tot} t_{mn}}{T_n^{AP}} \tag{4}$$

where $r_{mn}^{tot} = B_n \gamma_{mn}$ is the total achievable data rate (kbps) in AP n, γ_{mn} is the spectral efficiency (kbps/Hz) between MT m and AP n. The channel model used in [26] for Wi-Fi APs is adopted; $\gamma_{mn} = \log_2(1 + \frac{P_n g_{mn}}{\sigma^2})$ where P_n denotes the transmission power of AP n, σ^2 the noise power, and g_{mn} the channel gain between MT m and AP n encompassing the effects of path loss and antenna gains. It is assumed that APs operate on non-overlapping channels so that no interference exists among APs. In addition, since WLANs operate in an unlicensed band, APs do not interfere with BSs. Full power transmission in each time slot is assumed.

2.3. User-Centric Attributes

In this section, we present two user-centric attributes that are chosen to calculate the context-aware profit contributed upon associating MTs to networks. Both attributes have been considered in the literature as essential network selection parameters.

2.3.1. Signal quality

The signal quality is usually considered as an important attribute for making HO decisions in HWNs. However, it is difficult to compare the quality of the signal among different wireless access technologies because they have various maximum transmission power and receiver power thresholds. To overcome this issue, Shen *et al.* have proposed a signal quality formula in [27] that is applicable in different types of wireless technologies. The proposed formula is:

$$s_{mn} = \frac{P_{mn} - P_n^{th}}{P_n^{max} - P_n^{th}} \tag{5}$$

where P_n^{th} represents the receiver power threshold in network n, P_n^{max} the maximum transmitted signal power, and P_{mn} the actual signal power received by MT m. Shen *et al.* have managed to reduce their proposed formula to:

$$s_{mn} = 1 - \frac{\log(d_{mn})}{\log(R_n)} \tag{6}$$

Note that network n is unreachable by MT m if $d_{mn} > R_n$.

2.3.2. Instantaneous power consumption

MT power consumption is usually seen as an important user-centric attribute. Therefore, it is considered within the profit function. To estimate the instantaneous power consumed by MT m while receiving data from network n, the model empirically derived in [28] is used:

$$pc_{mn} = \alpha_n r_{mn} + \psi_n \tag{7}$$

where pc_{mn} is the power consumed by MT m while receiving data from network n, r_{mn} the downlink data rate in kbps, α_n (mW/kbps) and ψ_n (mW) are constants related to the wireless access technology of network n.

3. Optimization

3.1. Profit Function

In the proposed solution, users prefer to be served by a network with low instantaneous power consumption and high received signal quality. Consequently, a user-centric weighted profit function is defined to combine these two attributes. The weight of each attribute reflects its importance among other attributes in the profit function. These weights are set according to the user preferences. The weights of the signal quality and instantaneous power consumption for MT m are symbolized by w_m^s and w_m^{pc} respectively. Both weights are subject to the following constraint:

$$w_m^s + w_m^{pc} = 1 \tag{8}$$

Note that, both attributes (s_{mn} and pc_{mn}) have different measurement units. Thus, in order to be combined within the weighted profit function, a normalization step is required.

Table 1
Variable Definitions

Variable	Definition
K, M, N	Total number of SLs, MTs, and networks
$\mathcal{K}$, $\mathcal{M}$, $\mathcal{N}$	Sets of available SLs, MTs, and networks
R_n	Circular coverage radius of network n
l_m	Service level of MT m
ζ_n	Capacity of network n
β_{mn}	Weight of MT m in network n
$\varpi_n^{k^+}$	Weight of MTs with SL $> k$ in network n
s_{mn}	Quality of the signal received by MT m from network n
pc_{mn}	Power consumed by MT m while receiving data from network n
θ_k	Set of all MTs with SL k
w_m^s, w_m^{pc}	Preference weights for MT m related to the signal quality and power consumption
$\widehat{s}_{mn}$, $\widehat{pc}_{mn}$	Normalized s_{mn} and pc_{mn}
f_{mn}	Profit contributed upon associating MT m to network n
U_n	Total number of SBs in BS n
u_{mn}	Number of SBs that are allocated to MT m if it is connected to BS n
t_{mn}	Number of time slots allocated to MT m if it is connected to AP n
Q_m	Data rate requested by MT m
r_{mn}	The data rate that is supplied to MT m if it is connected to network n
r_{mn}^{tot}	The total data rate achievable by MT m with AP n
B_n^{RB}, B_n	Bandwidth of a RB and the total bandwidth of an AP respectively
γ_{mn}	Spectral efficiency between the BS or AP n and MT m
C_n	Total number of sub-channels in BS n
T_n^{AP}	Total number of time slots within one scheduling interval in AP n
T_n^{BS}	Total number of time slots within one scheduling interval in BS n

While normalizing, it is essential to differentiate between upward and downward attributes; attributes of which their higher value is preferable are called upward attributes; conversely, downward attributes are those we aim at decreasing their value. It is obvious that the signal quality is considered as an upward attribute while the instantaneous power consumption as a downward one. Therefore, based on [6], the normalized forms of the signal quality and instantaneous power consumption, respectively denoted by $\widehat{s}_{mn}$ and $\widehat{pc}_{mn}$, are:

$$\widehat{s}_{mn} = \frac{s_{mn}}{\max\limits_{m\in\theta_k, n\in\mathcal{N}}(s_{mn})} \tag{9}$$

$$\widehat{pc}_{mn} = \frac{1/pc_{mn}}{\max\limits_{m\in\theta_k, n\in\mathcal{N}}(1/pc_{mn})} \tag{10}$$

Note that increasing the value of $\widehat{s}_{mn}$ depends on increasing the value of s_{mn} while $\widehat{pc}_{mn}$ can be increased by decreasing pc_{mn}. Moreover, each attribute is normalized to the global maximum, *i.e.* $\max\limits_{m\in\theta_k, n\in\mathcal{N}}$, instead of the local one $(\max\limits_{n\in\mathcal{N}})$ because the profit function will be deployed in a global optimization problem.

The normalized profit function does not differentiate between MTs with unequal data rate requirements. Since the context of this paper considers MTs with different requested data rates, the profit function is multiplied by the amount of data rate requested by the MT to reflect the real profit contributed by each MT. Moreover, a MT m could connect to network n only if $R_n \geq d_{mn}$. Thus, a unit step function is defined as:

$$U(R_n - d_{mn}) = \begin{cases} 1, & \text{if } R_n \geq d_{mn}, \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

Hence, the overall profit of MT m in network n is:

$$f_{mn} = U(R_n - d_{mn}) \cdot \left(w_m^s \widehat{s}_{mn} + w_m^{pc} \widehat{pc}_{mn}\right) \cdot Q_m \tag{12}$$

Note that $\frac{f_{mn}}{Q_m}$ can be seen as the normalized profit of a MT, or the profit per kbps, which is the real profit contributed without multiplying by the requested data rate.

3.1.1. Profit-function-based network selection algorithm

In this section, we discuss the trivial profit-function-based network selection algorithm. Basically, the MT ranks candidate networks based on the profit function derived in Eq. (12). Normally, the MT targets the network with the highest profit value. Then, the MT estimates the number of resources that should be given by the target network in order to supply the MT with the requested data rate. If the targeted network have sufficient resources to serve the MT, the association between the MT and target network is established. Otherwise, the MT targets the network with the next higher rank, and so on, until the MT is associated to a network. The mechanism of estimating the number of requested resources is discussed throughout this paper, and the priority-aware profit-function-based network selection algorithm is discussed in Section 4.2.

3.1.2. Profit function characteristics

In this section, we highlight the characteristics of the profit function and its performance upon variating the weights w_m^s and w_m^{pc}. Therefore, three cases of the profit function are considered:

- Signal-quality-based profit function: $w_m^s = 1$ and $w_m^{pc} = 0$.
- Power-consumption-based profit function: $w_m^s = 0$ and $w_m^{pc} = 1$.
- Equal-weight-based profit function: $w_m^s = 0.5$ and $w_m^{pc} = 0.5$.

The performance of the profit function is evaluated in a scenario where each MT requests a specific number of resources, and each network has a limited amount of resources. Accordingly, we study the behavior of the profit-function-based network selection algorithm upon increasing the number of MTs in a heterogeneous wireless system that is based on Wi-Fi APs

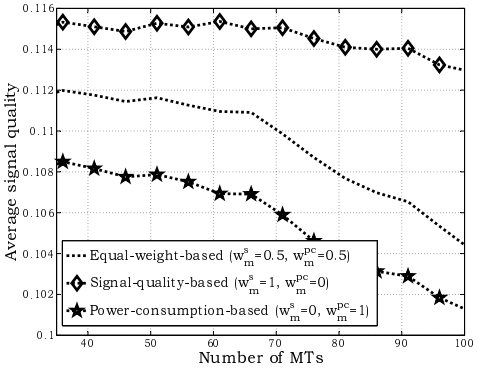

Figure 1: Average signal quality for different profit function cases. Based on Eq. (32).

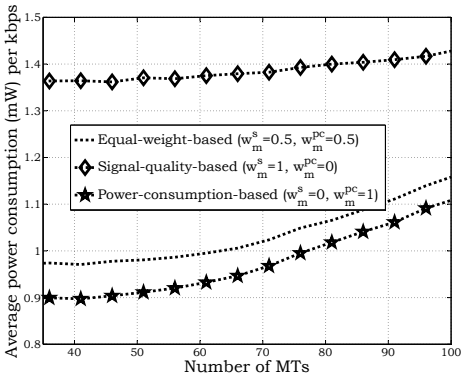

Figure 2: Average power consumption for different profit function cases. Based on Eq. (31).

and LTE BSs. Since we are only interested in showing the behaviour of the profit function, detailed simulation parameters are not discussed. However, they are slightly different than the parameters presented in Section 5.1 to ensure that MTs does not experience any blockage.

It is important to illustrate first that MTs aiming at only enhancing the signal quality, *i.e.* $w_m^s = 1$ and $w_m^{pc} = 0$, tends to attach to LTE BSs due to their long transmission range property. For example, when the number of MTs is 35, *i.e.* networks are not fully congested yet, 60% of the users aiming at only enhancing the signal quality are associated to LTE BSs. On the other hand, MTs aiming at only enhancing the power consumption, *i.e.* $w_m^s = 0$ and $w_m^{pc} = 1$, tends to associate to Wi-Fi APs due to their low power consumption property. It is noted that when the number of MTs is 35, 65% of MTs aiming at only enhancing the power consumption are connected to Wi-Fi APs.

It is shown in Fig. 1 that the signal-quality-based profit function maintains the highest signal quality, followed by the equal-weight-based profit function. The power-consumption-based profit function scores the lowest signal quality because $w_m^s = 0$. So, MTs tend to select Wi-Fi APs to save power, causing lower signal quality. As the number of MTs increases from 35 to 100, the average signal quality is decreased by only 2.1% for the signal-quality-based profit function, while the power-consumption-based profit function decreases the average signal quality by 6%. Increasing the number of MTs reduces the opportunity that MTs connect to their best available network. Therefore, the average signal quality decreases in general.

Since MTs following the signal-quality-based profit function tends to associate with LTE BSs, their average power consumption is high as shown in Fig. 2 because LTE networks requests higher power consumption.

On the other hand, MTs with the power-consumption-based perspective tends to associate with Wi-Fi APs due to the low power consumption property. Those MTs maintain the lowest power consumption. As the number of MTs increases, Wi-Fi APs become congested. Therefore, the average power consumption increases because MTs have less probability to be associated with their top-ranked network. It is remarkable that when the number of MTs is 40, the power-consumption-based profit function scores 34% less average power consumption than the signal-quality-based profit function. Therefore, the profit function responds explicitly to the variation of the weights in order to meet user preferences.

3.2. Optimization Problem

We aim at formulating an optimization problem to maximize the total profit for each SL. The problem should throw firm restrictions to prevent low-priority users from allocating resources that could be utilized by other users with higher priorities. A single network association should be ensured, as well as supplying the connected MT with data rate that is at least equal to its requested data rate threshold. Therefore, a set of $M \times N$ boolean user association variables x_{mn} are defined such that:

$$x_{mn} = \begin{cases} 1, & \text{if MT } m \text{ is associated to network } n, \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

Thus, the formulated problem is:

$$\textbf{P1: } max \quad \sum_{k\in\mathcal{K}} \sum_{m\in\theta_k} \sum_{n\in\mathcal{N}} f_{mn} x_{mn} \tag{14a}$$

$$s.t. \quad \sum_{m\in\theta_k} u_{mn} x_{mn} \leqslant U_n - \sum_{j>k} \sum_{i\in\theta_j} u_{in} x_{in} \qquad \forall\, k \in \mathcal{K},\ n \in \mathcal{N}_{\mathcal{BS}} \tag{14b}$$

$$\sum_{m\in\theta_k} t_{mn} x_{mn} \leqslant T_n^{AP} - \sum_{j>k} \sum_{i\in\theta_j} t_{in} x_{in} \qquad \forall\, k \in \mathcal{K},\ n \in \mathcal{N}_{\mathcal{AP}} \tag{14c}$$

$$\sum_{n\in\mathcal{N}} r_{mn} x_{mn} \geqslant \sum_{n\in\mathcal{N}} Q_m x_{mn} \qquad \forall m \in \mathcal{M} \tag{14d}$$

$$\sum_{n\in\mathcal{N}} x_{mn} \leqslant 1 \qquad \forall m \in \mathcal{M} \tag{14e}$$

$$x_{mn} \in \{0, 1\} \qquad \forall m \in \mathcal{M}, \forall n \in \mathcal{N} \tag{14f}$$

$$u_{mn} \in \mathbb{N}^+ \qquad \forall m \in \mathcal{M}, \forall n \in \mathcal{N}_{\mathcal{BS}} \tag{14g}$$

$$t_{mn} \in \mathbb{N}^+ \qquad \forall m \in \mathcal{M}, \forall n \in \mathcal{N}_{\mathcal{AP}} \tag{14h}$$

Constraint (14b) ensures that the capacity of LTE BSs is not exceeded and the resources allocated to high-priority MTs are not violated. Similarly, constraint (14c) guarantees the same aspects in Wi-Fi APs. Constraints (14e) and (14f) assure that a MT will be associated with a single network, or not connected at all (upon congestion). Constraint (14d) guarantees that the data rate received by a MT is at least equal to its requested data rate threshold. However, we are obliged to multiply both sides of the inequality by x_{mn} because upon congestion, some MTs will not be served. Constraint (14g) ensures that a single SB (LTE) is not assigned to multiple MTs simultaneously. Similarly, constraint (14h) guarantees that a single time slot in an AP is not allocated for multiple MTs at the same time. Note that MTs are distributed in different SL sets θ_k, and constraints (14b) and (14c) ensure that the resources allocated for MTs with SLs higher than k, *i.e.* MTs $\in \theta_j$ such that $j > k$, are not given to MTs with SL k, *i.e.* MTs $\in \theta_k$. Therefore, it is preferable to show the maximization form in terms of all SLs and all MTs in SL sets instead of directly maximizing for all MTs, *i.e.* "$max \sum_{k\in\mathcal{K}}\sum_{m\in\theta_k}\sum_{n\in\mathcal{N}}$" instead of "$max \sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}}$". This plays a role in clarifying the characteristics of the formulated problem.

3.3. Problem Simplification

The formulated problem (**P1**) aims at finding three sets of variables:

- The boolean association variables (x_{mn}).
- The number of SBs allocated for each MT m connected to BS n (u_{mn}).
- The number of time slots allocated for each MT m connected to AP n (t_{mn}).

In the following, the number of resources that should be allocated by each network in order to supply the MT with its requested data rate (if the MT is associated to the network) is calculated. Hence, u_{mn} or t_{mn} can be seen as the weight of MT m in network n. Thus, the optimization problem now aims at finding only the boolean association variables x_{mn}. Therefore, based on constraint (14d), we calculate the number of resources that should be allocated by each network to supply MTs with their minimum requested data rate, *i.e.* $r_{mn} = Q_m \; \forall m \in \mathcal{M}$. Hence, based on (3), and according to the approach adopted by [19] and [29], the minimum number of resources that should be allocated to MT m if it is connected to BS n is:

$$u_{mn} = \frac{Q_m T_n^{BS}}{B_n^{RB}\gamma_{mn}} \qquad \forall n \in \mathcal{N}_{\mathcal{BS}} \quad (15)$$

Similarly, and based on (4) for Wi-Fi APs:

$$t_{mn} = \frac{Q_m T_n^{AP}}{r_{mn}^{tot}} \qquad \forall n \in \mathcal{N}_{\mathcal{AP}} \quad (16)$$

In fact, both (15) and (16) can be seen as the weights of MTs in networks. Thus, $\beta_{mn} \in \mathbb{N}^+$ is introduced to indicate the weight of MT m in network $n \in \mathcal{N}$ such that:

$$\beta_{mn} = \begin{cases} \left\lceil \frac{Q_m T_n^{BS}}{B_n^{RB}\gamma_{mn}} \right\rceil & \forall n \in \mathcal{N}_{\mathcal{BS}} \\ \left\lceil \frac{Q_m T_n^{AP}}{r_{mn}^{tot}} \right\rceil & \forall n \in \mathcal{N}_{\mathcal{AP}} \end{cases} \quad (17)$$

The ceiling ($\lceil . \rceil$) of values in (15) and (16) is taken to preserve the integral constraints (14g) and (14h). Similarly, ζ_n denotes the capacity of network $n \in \mathcal{N}$ such that:

$$\zeta_n = \begin{cases} U_n & \forall n \in \mathcal{N}_{\mathcal{BS}} \\ T_n^{AP} & \forall n \in \mathcal{N}_{\mathcal{AP}} \end{cases} \quad (18)$$

Therefore, based on (17) and (18), **P1** could be reformulated as:

$$\textbf{P2: } max \sum_{k\in\mathcal{K}}\sum_{m\in\theta_k}\sum_{n\in\mathcal{N}} f_{mn}x_{mn} \quad (19a)$$

$$s.t. \sum_{m\in\theta_k}\beta_{mn}x_{mn} \leqslant \zeta_n - \sum_{j>k}\sum_{i\in\theta_j}\beta_{in}x_{in} \qquad \forall\, k \in \mathcal{K},\; n \in \mathcal{N} \quad (19b)$$

$$\sum_{n\in\mathcal{N}} x_{mn} \leqslant 1 \qquad \forall m \in \mathcal{M} \quad (19c)$$

$$x_{mn} \in \{0,1\} \qquad \forall m \in \mathcal{M}, \forall n \in \mathcal{N} \quad (19d)$$

Actually **P2** could be further simplified by fixing the value $\sum_{j>k}\sum_{i\in\theta_j}\beta_{in}x_{in}$ in constraint (19b). To do so, a new variable $\varpi_n^{k^+}$ is introduced to express the number of resources that are allocated to MTs with SL $> k$ in network n, *i.e.* the total weight of MTs with SL $> k$. Thus:

$$\varpi_n^{k^+} = \begin{cases} 0 & if\ k = K \\ \sum_{j>k}\sum_{i\in\theta_j}\beta_{in}x_{in} & if\ k < K \end{cases} \quad (20)$$

Note that $\varpi_n^{k^+}$ depends on the association results of MTs with SL $> k$. Hence, if the association decision for MTs with SL $> k$ is found, $\varpi_n^{k^+}$ can be considered as a constant value for MTs with SL k. Therefore, **P2** is distributed to K problems which will be solved sequentially according to the decreased order of priority, *i.e.* $K, K-1, \ldots, 1$. Thus, the user association and resource allocation problem for MTs with SL k is:

$$\textbf{P3: } max \sum_{m\in\theta_k}\sum_{n\in\mathcal{N}} f_{mn}x_{mn} \quad (21a)$$

$$s.t. \sum_{m\in\theta_k}\beta_{mn}x_{mn} \leqslant \zeta_n - \varpi_n^{k^+} \qquad \forall n \in \mathcal{N} \quad (21b)$$

$$\sum_{n\in\mathcal{N}} x_{mn} \leqslant 1 \qquad \forall m \in \theta_k \quad (21c)$$

$$x_{mn} \in \{0,1\} \qquad \forall m \in \theta_k, \forall n \in \mathcal{N} \quad (21d)$$

The formulated problem **P3** consists of finding an optimal set of association variables from a finite set of objects. In such problems, exhaustive search is not feasible for even a small-sized

systems. Therefore, solving the problem is not straightforward. **P3** operates on the domain of optimization problems where the set of feasible solutions is discrete, and in which the target is to find the best solution. Hence, the complexity of finding the optimal solution is $O(N^{|\theta_k|})$ where |.| denotes the cardinality of a set. All the discussed complexities are listed in Table 4. The variables in **P3** are restricted to have binary values and the objective function and constraints are linear, thus it is considered as BLP. Therefore, classical approaches used to solve continuous optimization problems could not be deployed to solve problem **P3**.

In fact, BLP, usually referred to as 0-1 integer linear programming problem, is one of the Karp's 21 NP-complete problems [30]. One class of algorithms used to solve BLPs are variants of the branch and bound method. To evaluate the optimal solution based on the branch and bound algorithm, the GNU linear programming kit (GLPK) is used. GLPK is intended to solve integer and linear programming optimization problems. However, as the number of variables grows largely, the optimal solution becomes intractable. Therefore, in this paper, a solution with polynomial-time complexity is proposed to approximate problem **P3**. Note that the linearity of the problem is explicitly discussed in the next section.

3.4. Relaxation of the Binary Constraint

In this section, the continuous relaxation approach is considered to deal with the binary constraint (21d). Accordingly, the binary association constraint (21d) is relaxed to a continuous. Thus, each MT is now allowed to access multiple networks simultaneously, *i.e.* multi-homing. Then, a new methodology is proposed in Algorithm 1 to preserve constraint (21d) by considering only boolean association results. Relaxing the binary constraint permits solving the optimization problem using standard mathematical methods that can solve linear programs. The relaxed problem is:

$$\textbf{P4: } max \sum_{m\in\theta_k}\sum_{n\in\mathcal{N}} f_{mn}x_{mn} \tag{22a}$$

$$s.t. \sum_{m\in\theta_k} \beta_{mn}x_{mn} \leqslant \zeta_n - \varpi_n^{k^+} \qquad \forall n \in \mathcal{N} \tag{22b}$$

$$\sum_{n\in\mathcal{N}} x_{mn} \leqslant 1 \qquad \forall m \in \theta_k \tag{22c}$$

$$0 \leq x_{mn} \leq 1 \qquad \forall m \in \theta_k, \forall n \in \mathcal{N} \tag{22d}$$

where the inequalities (22b), (22c), and (22d) specify a convex polytope over which the profit function is to be optimized. It is essential to present the methodology of expressing problem **P4** in the standard form of a linear program. Since the binary constraint is relaxed, constraint (22d) could be replaced now by $x_{mn} \geq 0$ because the set of constraints (22c) ensure that $x_{mn} \leq 1\ \forall m, n$. Hence, problem **P4** could be expressed in the standard canonical form of a linear program such that:

$$\begin{aligned} max \quad & \boldsymbol{c}^\top \boldsymbol{x} \\ s.t. \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \\ & \boldsymbol{x} \geq 0 \end{aligned} \tag{23}$$

where:

- $\boldsymbol{c} \in \mathbb{R}^{|\theta_k|N}$ is a vector that contains all the profit values f_{mn}, and $(\cdot)^\top$ is the matrix transpose.
- $\boldsymbol{x} \in \mathbb{R}^{|\theta_k|N}$ is a vector that contains all the user association variables x_{mn}.
- $\boldsymbol{A} \in \mathbb{R}^{(|\theta_k|+N)\times(|\theta_k|N)}$ and $\boldsymbol{b} \in \mathbb{R}^{|\theta_k|+N}$ are respectively a matrix and a vector of coefficients related to constraints (22b) and (22c); $\boldsymbol{A}$ contains the coefficients at the left side of the inequalities in the constraints, and $\boldsymbol{b}$ the constants on the right side.

For example, let us consider a heterogeneous wireless system with three MTs of SL k placed within the overlapped coverage range of two networks, then:

$$\boldsymbol{c}^\top = \begin{bmatrix} f_{11} & f_{12} & f_{21} & f_{22} & f_{31} & f_{32} \end{bmatrix},$$
$$\boldsymbol{x} = \begin{bmatrix} x_{11} & x_{12} & x_{21} & x_{22} & x_{31} & x_{32} \end{bmatrix}^\top,$$

$$\boldsymbol{A} = \overbrace{\begin{bmatrix} \beta_{11} & 0 & \beta_{21} & 0 & \beta_{31} & 0 \\ 0 & \beta_{12} & 0 & \beta_{22} & 0 & \beta_{32} \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}}^{|\theta_k|N} \begin{matrix} \left.\begin{matrix} \\ \\ \end{matrix}\right\} N \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} |\theta_k| \end{matrix} \begin{matrix} \left.\begin{matrix} \\ \\ \end{matrix}\right\} (22b) \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} (22c) \end{matrix}, \boldsymbol{b} = \begin{bmatrix} \zeta_1 - \varpi_1^{k^+} \\ \zeta_2 - \varpi_2^{k^+} \\ 1 \\ 1 \\ 1 \end{bmatrix} \tag{24}$$

Note that problem **P4** can be seen now as a standard linear programming problem. Expressing problem **P4** in the form shown in (23) also serves as a proof of linearity for problems **P3** and **P4**.

The number of variables in the standard linear program **P4** is $|\theta_k|N$. In practice, the simplex method performs very well when used to solve this linear program even for a large number of variables. However, its worst-case computational complexity is exponential [31]. Other methods with polynomial-time complexity have been proposed to solve standard linear programs. The interior-point methods are preferred among them; the theoretical computational complexity is $O(|\theta_k|^3 N^3 L)$, where L is the length of the binary coding of the input data [31]. The fact that the complexity depends on L implies that the time required to solve the problem increases with the required accuracy of the computations.

Solving problem **P4** results in three sets of MTs classified according to their association status:

- $\mathcal{S}_1^k$: set of associated MTs with boolean association values.
- $\mathcal{S}_2^k$: set of associated MTs with fractional association values.
- $\mathcal{S}_3^k$: set of unassociated MTs.

Algorithm 1: Relaxation-based solution

Output: Association variables for all MTs in θ_k

1.1: **foreach** $n \in \mathcal{N}$ **do**
1.2: $\quad \overline{\zeta_n} := \zeta_n - \varpi_n^{k^+}$;
1.3: **end**
1.4: $\overline{\theta_k} := \theta_k$;
1.5: **while** $\overline{\theta_k} \neq \phi$ **do**
1.6: $\quad$ Solve problem **P4** $\forall m \in \overline{\theta_k}$ and according to $\overline{\zeta_n}$;
1.7: $\quad \mathcal{S}_1^k := \{m \in \overline{\mathcal{M}} : \sum_{n \in \mathcal{N}} \lfloor x_{mj} \rfloor = 1\}$;
1.8: $\quad \mathcal{S}_3^k := \{m \in \overline{\mathcal{M}} : \sum_{n \in \mathcal{N}} x_{mj} = 0\}$;
1.9: $\quad \mathcal{S}_2^k := \overline{\theta_k} - \mathcal{S}_1^k - \mathcal{S}_3^k$;
1.10: $\quad$ **if** $|\mathcal{S}_2^k| = \phi$ **then**
1.11: $\quad\quad$ Save the association values $\forall m \in \mathcal{S}_1^k \cup \mathcal{S}_3^k$;
1.12: $\quad\quad \overline{\theta_k} = \phi$;
1.13: $\quad$ **else**
1.14: $\quad\quad$ Save the association values $\forall m \in \mathcal{S}_1^k$;
1.15: $\quad\quad \overline{\theta_k} = \overline{\theta_k} - \mathcal{S}_1^k$;
1.16: $\quad\quad$ **foreach** $n \in \mathcal{N}$ **do**
1.17: $\quad\quad\quad \overline{\zeta_n} = \overline{\zeta_n} - \sum_{m \in \mathcal{S}_1^k} \beta_{mn} x_{mn}$;
1.18: $\quad\quad$ **end**
1.19: $\quad$ **end**
1.20: **end**

Since in our context, a MT could be associated with a single network while receiving its requested data rate, a new approach is used to assign MTs to appropriate networks and empty the set $\mathcal{S}_2^k$. Algorithm 1 shows our proposed method.

Algorithm 1 keeps solving the relaxed problem **P4** for all MTs in $\mathcal{S}_1^k \cup \mathcal{S}_3^k$ until $|\mathcal{S}_2^k| = 0$, *i.e.* all the results of problem **P4** are binary. Every time the optimization problem is solved, the association values of MTs in $\mathcal{S}_1^k$ are saved, the number of free resources in each network is updated (*lines 1.16-1.18*), and MTs in $\mathcal{S}_1^k$ are not considered within the optimization function anymore (*line 1.15*).

3.5. The Proposed Approximation-based Solution

After taking a closer look at problem **P3**, we notice that it is similar to the generalized assignment problem (GAP) [32]. In fact, Martello and Toth, who have significant contributions in the domain of GAP, knapsack, and bin-packing problems, have proposed a heuristic algorithm to approximate GAP based on an ordering of the MTs [32]. There, the "desirability" of assigning MT m to network n is measured according to a desirability factor Ω_{mn}. The possible factors that could be considered as a desirability measure are discussed in Section 3.5.1. For each MT, the difference between the highest and the second highest value of Ω_{mn} is computed, and MTs are then assigned in the decreasing order of this difference. That is, each MT is assigned to its best network according to the following criteria:

$$\max_{n} \min_{n \neq n\prime} (\Omega_{mn\prime} - \Omega_{mn}) \tag{25}$$

or in other words:

$$\begin{aligned} &\min_{n \neq n\prime} \Omega_{mn\prime} - \Omega_{mn} \\ &\text{where} \\ &n\prime = \arg\max_{n} \Omega_{mn} \end{aligned} \tag{26}$$

The computational experiments conducted by Martello and Toth have shown that good results are obtained using this algorithm. However, their proposed algorithm does not exactly suit problem **P3** for two reasons:

- The algorithm is designed to solve GAP while constraint (21c) is replaced by $\sum_{n \in \mathcal{N}} x_{mn} = 1$. That is, all MTs should be associated to networks. While upon congestion, some MTs would not be able to associate to any network. Then, the algorithm would fail to approximate problem **P3**.
- The algorithm assumes that all networks are reachable by all MTs. Therefore, it does not differentiate between MTs reachable by a single network and others reachable by multiple networks.

Thus, we modify their proposed algorithm to adapt problem **P3** as shown in Algorithm 2.

At first, all association variables for MTs in θ_k are set to 0. Algorithm 2 iteratively considers all the unassociated MTs, and determines the MT m^* having the maximum difference between the highest and the second highest Ω_{mn} ($n \in F_m$ where F_m is defined in *line 2.9*). MT m^* is then assigned to the network for which Ω_{m^*n} is maximum, *i.e.* network n^*. It is this property of the algorithm which leads to significant results when tested; the algorithm considers the second maximum Ω_{mn} instead of focusing only on the first maximum. Moreover, after taking each association decision, the algorithm re-evaluates, for each MT, the maximum difference between the highest and the second highest Ω_{mn}, and associates MTs based on these new results. Thus, a semi-global view on the available networks and their profit is maintained while taking association decisions. In addition, the algorithm prefers to first associate MTs with only one available network, *i.e.* $|F_m| = 1$. We add the if block in *lines 2.16-2.20* to associate the MT with highest Ω_{mn} among other MTs with a single available network. Initially, the original algorithm associates any MT with a single available network without taking into consideration the value of Ω_{mn}. This aspect of the algorithm plays a vital role in decreasing the blocking probability.

Algorithm 2 can be implemented efficiently by initially sorting in decreased order, for each MT m, the values Ω_{mn} ($n \in \mathcal{N}$). This requires $O(N \log N)$ for a single MT. Thus for all MTs $m \in \theta_k$ it requires $O(|\theta_k| N \log N)$. The sorting step makes immediately available, at each iteration in the inner loop, the pointers to the maximum and the second maximum Ω_{mn}. Hence, the main while loop performs the $O(|\theta_k|)$ associations within a total of $O(|\theta_k|^2)$ iterations; whenever a MT is assigned, the decrease in $\overline{\zeta_{n^*}}$ makes it necessary to update the pointers. Since, however, the above maxima can only decrease during execution, a total of $O(|\theta_k|^2)$ operations is required for these checks and updates. Thus, we conclude that the overall complexity of Algorithm 2 is $O(|\theta_k|^2 + |\theta_k| N \log N)$.

Algorithm 2: Approximation algorithm

2.1: $\mathcal{U} := \theta_k$;
2.2: **foreach** $n \in \mathcal{N}$ **do**
2.3: $\quad \overline{\zeta_n} := \zeta_n - \varpi_n^{k^+}$;
2.4: **end**
2.5: **while** $\mathcal{U} \neq \phi$ **do**
2.6: $\quad c^* := -\infty$;
2.7: $\quad d^* := -\infty$;
2.8: $\quad$ **foreach** $m \in \mathcal{U}$ **do**
2.9: $\quad\quad F_m := \{n \in \mathcal{N} : f_{mn} \neq 0 \wedge \beta_{mn} \leq \overline{\zeta_n}\}$;
2.10: $\quad\quad$ **if** $F_m = \phi$ **then**
2.11: $\quad\quad\quad \mathcal{U} = \mathcal{U} - \{m\}$;
2.12: $\quad\quad$ **else**
2.13: $\quad\quad\quad n\prime = argmax_n\{\Omega_{mn} : n \in F_m\}$;
2.14: $\quad\quad\quad$ **if** $|F_m|$=1 **then**
2.15: $\quad\quad\quad\quad d := +\infty$;
2.16: $\quad\quad\quad\quad$ **if** $c^* < \Omega_{mn\prime}$ **then**
2.17: $\quad\quad\quad\quad\quad c^* = \Omega_{mn\prime}$;
2.18: $\quad\quad\quad\quad\quad n^* := n\prime$;
2.19: $\quad\quad\quad\quad\quad m^* := m$;
2.20: $\quad\quad\quad\quad$ **end**
2.21: $\quad\quad\quad$ **else**
2.22: $\quad\quad\quad\quad d := \Omega_{mn\prime} - max_2\{\Omega_{mn} : n \in F_m\}$;
2.23: $\quad\quad\quad\quad$ **if** $d > d^*$ **then**
2.24: $\quad\quad\quad\quad\quad d^* = d$;
2.25: $\quad\quad\quad\quad\quad n^* := n\prime$;
2.26: $\quad\quad\quad\quad\quad m^* := m$;
2.27: $\quad\quad\quad\quad$ **end**
2.28: $\quad\quad\quad$ **end**
2.29: $\quad\quad$ **end**
2.30: $\quad$ **end**
2.31: $\quad$ **if** $d \neq -\infty$ **then**
2.32: $\quad\quad x_{m^*n^*} = 1$;
2.33: $\quad\quad \mathcal{U} = \mathcal{U} - \{m^*\}$;
2.34: $\quad\quad \overline{\zeta_{n^*}} = \overline{\zeta_{n^*}} - \beta_{m^*n^*}$;
2.35: $\quad$ **end**
2.36: **end**

Algorithm 3: Greedy algorithm

Output: Association variables for all MTs in θ_k
3.1: **foreach** $n \in \mathcal{N}$ **do**
3.2: $\quad \overline{\zeta_n} := \zeta_n - \varpi_n^{k^+}$;
3.3: **end**
3.4: $\mathcal{X} := \{\Omega_{mn} : m \in \theta_k \wedge f_{mn} \neq 0\}$;
3.5: **while** $\mathcal{X} \neq \phi$ **do**
3.6: $\quad m\prime = argmax_m(\mathcal{X})$;
3.7: $\quad n\prime = argmax_n(\mathcal{X})$;
3.8: $\quad$ **if** $\sum_{n\in\mathcal{N}}\lfloor x_{m\prime n}\rfloor = 0$ **then**
3.9: $\quad\quad$ **if** $\beta_{m\prime n\prime} \leq \overline{\zeta_{n\prime}}$ **then**
3.10: $\quad\quad\quad x_{m\prime n\prime} = 1$;
3.11: $\quad\quad\quad \overline{\zeta_{n\prime}} = \overline{\zeta_{n\prime}} - \beta_{m\prime n\prime}$;
3.12: $\quad\quad$ **else**
3.13: $\quad\quad\quad x_{m\prime n\prime} = 0$;
3.14: $\quad\quad$ **end**
3.15: $\quad$ **else**
3.16: $\quad\quad x_{m\prime n\prime} = 0$;
3.17: $\quad$ **end**
3.18: $\quad \mathcal{X} = \mathcal{X} - \{\Omega_{m\prime n\prime}\}$;
3.19: **end**

3.5.1. Efficiency factor

Problem **P3** aims at maximizing the profit in the system. Therefore, Martello and Toth have proposed in [32] to use the profit (f_{mn}) or $\frac{\text{profit}}{\text{weight}}$ as desirability factor in Algorithm 2. Since problem **P3** deals with MTs having different data rate requirements, *i.e.* different weights, the $\frac{\text{profit}}{\text{weight}}$ is suitable as desirability factor for this problem. However, considering the weight of MTs, which can be seen as the number of requested resources, is not straightforward because access technologies have different types and amounts of resources. As a matter of fact, the amount of bandwidth that should be supplied by a network to a MT is related to the channel conditions between the MT and the network, and to the amount of data rate requested by the MT. Moreover, the bandwidth (in Hz) is a limited resource in all communication systems. Therefore, the amount of bandwidth requested by a MT from a network could be considered as a weight. The amount of bandwidth requested by MT m from BS n is $\frac{B_n^{RB} u_{mn}}{T_n^{BS}}$, and from AP n is $\frac{B_n t_{mn}}{T_n^{AP}}$. Hence, the efficiency e_{mn} is introduced to denote the profit per weight (requested bandwidth) contributed to the system upon associating MT m to network n such that:

$$e_{mn} = \begin{cases} \frac{f_{mn}}{B_n^{RB}(u_{mn}/T_n^{BS})} & \forall n \in \mathcal{N}_{\mathcal{BS}} \\ \\ \frac{f_{mn}}{B_n(t_{mn}/T_n^{AP})} & \forall n \in \mathcal{N}_{\mathcal{AP}} \end{cases} \tag{27}$$

The normalized profit ($\frac{f_{mn}}{Q_m}$) could be also used as desirability factor, but it does not consider the number of requested resources. Although the normalized profit reflects the actual profit contributed upon associating a MT, however, the channel quality between the MT and the BS or AP is not considered. Therefore, the efficiency is chosen as a main desirability factor. The difference between using the efficiency (e_{mn}) and the normalized profit ($\frac{f_{mn}}{Q_m}$) as desirability factors is discussed in Section 5.3.5.

3.5.2. Simple greedy solution

To explore the importance of assigning MTs based on the criteria proposed in (25), or (26), we would like to compare the performance of Algorithm 2 to a simpler greedy heuristic solution that orders MTs based on their direct maximum desirability value only, *i.e.* $\max_{\forall n\in\mathcal{N}} \Omega_{mn}$. Therefore, we consider the greedy solution shown in Algorithm 3. All the desirability values Ω_{mn} are sorted in decreasing order in set $\mathcal{X}$. In each iteration, the unassociated MT with the highest desirability measure, *i.e.* MT $m\prime$, is associated if its target network $n\prime$ has sufficient resources.

Sorting the desirability values in $\mathcal{X}$ in descending order makes the pointer for the maximum value immediately available in each iteration. Since the number of variables is $|\theta_k|N$, then the sorting complexity is $O(|\theta_k|N \log |\theta_k|N)$. Hence, the

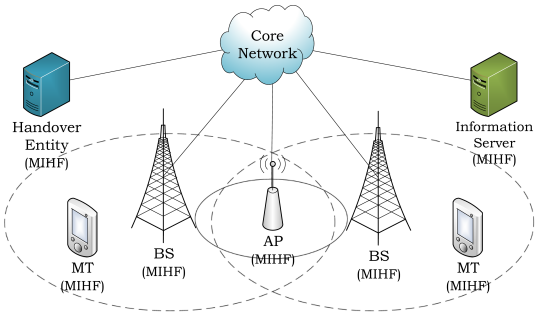

Figure 3: IEEE 802.21-based heterogeneous wireless system with a centralized HOE responsible for making user association and resource allocation decesions.

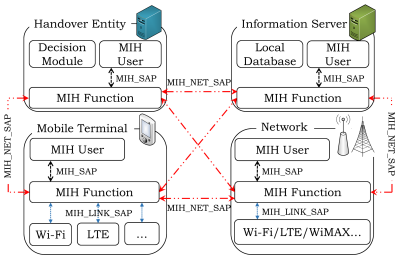

Figure 4: Communication interfaces between different IEEE 802.21 MIH layers and entities (local or remote) through MIH SAPs.

algorithm's complexity is $O(|\theta_k|N + |\theta_k|N \log |\theta_k|N)$, because it requires $|\theta_k|N$ iterations to iterate all the variables.

4. System Architecture and Solution Management Strategy

When an active connection is handed off between networks that are of the same access technology, the HO can usually be executed within that access technology itself. For example, a VoIP call over Wi-Fi can be handed over between APs using Wi-Fi standards such as 802.11f and 802.11r. However, if it is required to perform HO between two networks of different access technologies, *e.g.* from Wi-Fi AP to LTE BS, then an external protocol is required to manage the HO. In 2008, IEEE has published a new standard which is the 802.21 media-independent HO (MIH) [33] to enable seamless HO between networks of same or different types. MIH can communicate with several network protocols to facilitate the HO procedures. Those protocols include the session initiation protocol (SIP) for signaling and mobile IP protocol for mobility management. The standard is intended for HWNs integrating both 802 and non-802 access technologies.

4.1. System Architecture

In this paper, we rely on a centralized HO entity (HOE) to manage user association and allocate the downlink resources of HWNs. To fulfill its purpose, the HOE is empowered with the 802.21 MIH capabilities. The IEEE 802.21 standard defines an MIH framework that is intended to optimize the HO process in HWNs. The standard equips ordinary network entities, shown in Fig. 3, with MIH functionalities (MIHF) to facilitate the HO decision, coordinate user association, and allocate network resources. Moreover, the MIH standard supplies a common platform for exchanging contextual information which could be classified into user-centric, service-centric, and network-centric context. User-centric, as mentioned earlier, determines the user preferences, power consumption, and signal quality. The service-centric is related to the number of resources requested by each MT. The network-centric context is based on the instantaneously available resources at networks, their geographical location, type, and characteristics.

The IEEE 802.21 also defines the information server (IS) as an MIH entity with a local database that contains static and dynamic information about users and networks. According to the standard several regional ISs might exist.

The MIHF allows a higher layer in network entities, referred to as MIH User (Fig. 4), to interact with the lower link layer while the access technology of the latter is completely abstracted. The communication interface between remote MIHFs, *i.e.* MIHFs on remote entities, or between the MIHF and other layers in the same (local) entity, is based on a number of defined service primitives that are grouped in service access points (SAPs). Fig. 4 shows the interaction between different entities and layers according to the following SAPs:

- MIH_SAP: media-independent SAP that allows communication between the MIHF layer and the higher-layer MIH User. It also provides an interface for MIH Users to control and monitor different links regardless of their access technology.
- MIH_LINK_SAP: media-dependent SAP that acts as an interface between the local MIHF and the lower link layer.
- MIH_NET_SAP: media-dependent SAP that provides transport services over the data plane enabling message exchange between remote MIHFs. For instance, this SAP is used to exchange messages between MTs, HOE, and IS.

The proposed centralized HOE maintains a global view on the system and communicates with different entities through the MIH protocol. The HOE consists of MIH User, decision module, and MIHF (Fig. 4). The decision module is responsible for taking user association and resource allocation decisions. These decisions are based on information collected from the system. Specifically, the local MIHF on the HOE communicates with the remote MIHFs of networks, ISs, and MTs. The decision module requests those information from MIHF using the MIH User. The abstraction of the MIH User layer and the media-independent SAPs allows the same procedures (functions) to run on different access technologies. Thus, any future technology could be supported by the MIH protocol after defining its media-dependent SAPs.

MIHF encompasses different types of services to assist the HO process and exchange messages between different entities.

Those services are of three types:

- Media-independent event service (MIES): detects and reports changes in the physical, data link, and logical link layers. For example, it can report that the spectral efficiency has degraded below a certain threshold. Events can be reported to local and/or remote MIH Users.
- Media-independent command service (MICS): provides a set of commands to control the link layer state. Commands can be invoked by local or remote MIH Users. For instance, "INITIATE_HO" is a command in which the MIHF of the HOE provides to the MT's MIHF. This command includes the ID, or SSID, of an alternative BS or AP that the MT could use. Moreover, it could be used to set spectral efficiency thresholds.
- Media-independent information service (MIIS): provides a framework for MIH entities to collect static and dynamic information useful for making HO decisions. Information can be related to MT's requested data rate and QoS, geographic location of networks and MTs, link layer address, the capacity of networks, *etc.*

4.2. Solution Management Strategy

In this section, we discuss the proposed solution management strategy where a MT with low priority is not allowed to utilize resources allocated for MTs with higher priorities. The solution management strategy tries to minimize the number of times the optimization function is processed without affecting the optimality of the algorithm. Moreover, the aspects that trigger the resource allocation and user association algorithm are discussed. Mainly, the solution management strategy tries to decrease the number of MTs that are involved within the optimization problem (**P3**). The resource allocation and user association algorithm is triggered when one of the following scenarios occurs:

- The current serving network is not able to supply a certain MT with its requested data rate.
- A new connection is initiated.
- A MT with an active connection is about to leave the boundaries of its serving network.

However, it is not always required to run the optimization function. For instance, if a new connection is initiated, the MT could evaluate its candidate networks, and try to connect to the best one. If the target network has sufficient resources to serve the newly admitted connection, the MT will connect without having to run the optimization function. On the other hand, the target network might not be able to serve the MT unless it dissociates some MTs with lower SL. In this case, lets assume that the newly admitted MT has a SL of 3, it might be enough to dissociate some MTs of SL 1, *i.e.* run the optimization function for MTs with SL 1, without having to encompass MTs with other SLs within the optimization problem. The fact that we have distributed problem **P2** into K problems **P3**, each for a specific SL, enables applying such strategy without violating the optimality of the solution.

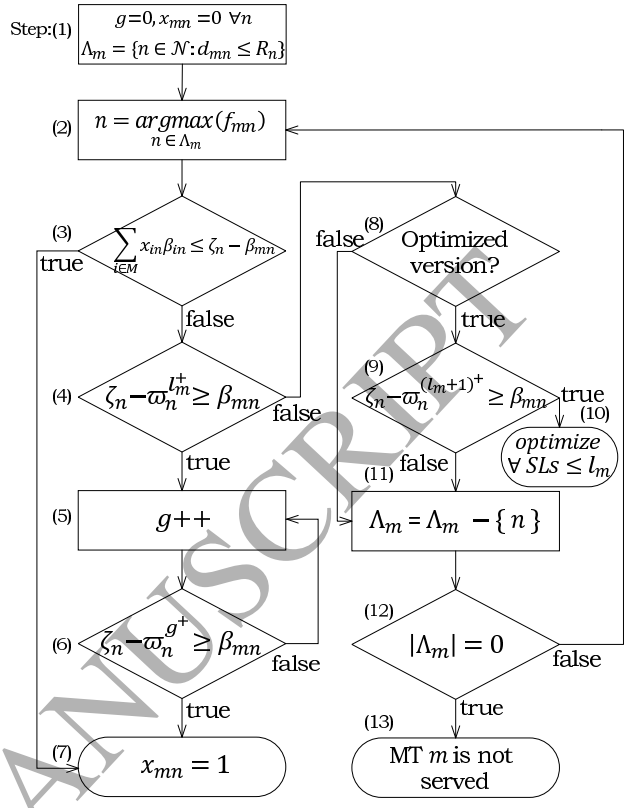

Figure 5: The proposed algorithm that determines the association values of MT m when it experience one of the scenarios that trigger the resource allocation and MT association algorithm.

MT m undergoes the procedures shown in Fig. 5 upon experiencing any of the scenarios that trigger the resource allocation and user association algorithm. The flow chart outputs the association variables x_{mn} for MT m, and an integer value g where all MTs with SLs $< g$ undergo the same procedures (shown in Fig. 5), as well as some or all MTs with SL g. The optimized version of the solution (Fig. 5-*step 8*) indicates using one of the methods proposed to solve or approximate problem **P3**. When the optimized version of the solution is deployed, the algorithm in Fig. 5 finds the minimal number of SLs that will undergo the optimization problem **P3**. On the contrary, if the optimized version is not deployed, the algorithm describes the profit-function-based solution for the problem.

The algorithm in Fig. 5 is detailed as follows: *Step 1* is an initialization step where integer g and association variables x_{mn} ($n\in\mathcal{N}$) are set to zero. Λ_m denotes the set of all networks for which MT m is within their coverage range. In *step 2*, the algorithm chooses the network n having the highest profit. The algorithm tests in *step 3* if the unallocated bandwidth resources of the selected network are sufficient to serve MT m. If that is the case, x_{mn} is immediately set to 1 (*step 7*), which indicates the association of MT m to network n. Conversely, if the number of requested resources is more than the unallocated ones, the algorithm proceeds to the next step. The association value

x_{mn} could be immediately determined in *step 4*. If the number of resources that are not allocated to MTs with SLs $\geq l_m$ is sufficient to serve MT m, then this MT will be surely associated to network n. However, the algorithm enters an iterative process in *step 5* and *step 6* to determine the SL(s) of MTs that might be detached from the selected network. Of course, it is preferable to detach MTs of the lowest SL first. Hence, g increases by each iteration. On the other hand, in *step 4*, if the number of resources that are not allocated to MTs with SLs $\geq l_m$ is less than the number of resources requested by MT m, and if the optimized version of the solution is not deployed (*step 8*), then the algorithm tries to associate MT m to the next top-ranked network (*i.e.* the network with second highest profit in Λ_m). To do so, the selected network is removed from the list of available networks in *step 11*. *Step 12* tests if the cardinality of the available networks set is equal to 0, which indicates that the algorithm has already tried to associate MT m to all its reachable networks. If so, MT m will not be served as indicated in *step 13*. Otherwise, the algorithm tries to associate MT m to its next top-ranked network.

On the other hand, upon congestion, the optimized version of the solution allows MT m to use resources allocated for MTs with SLs $\leq l_m$. In the profit-function-based solution, MT m is not allowed to allocate resources utilized by MTs with SL $= l_m$. The idea here is to maximize the profit of MTs with SLs $\leq l_m$ by efficiently utilizing the resources. *Step 9* mainly tests if the number of resources that are not allocated to MTs with SLs $> l_m$ is sufficient to serve MT m. In this case, the optimization problem **P3** is sequentially processed, in the decreasing order of SL, for each SL $\leq l_m$ (*step 10*).

4.3. User Priority Assignment

Throughout this paper, we have discussed the user association and resource allocation problem in HWNs with users having different priorities. However, the aspects that should be considered upon assigning user priorities are not discussed. Therefore, two general scenarios are presented.

The first scenario is based on a service level agreement (SLA) that could be signed between the user and the system operator. The system operator provides several SLs, each having a different pricing plan. Of course, it is expected that the best SL will have the most expensive pricing plan. Users are assigned to SLs according to their selected pricing scheme and the amount of money they are willing to pay in order to experience better service. The lowest SL is assigned for users who are not willing to pay extra money in order to experience better service.

The second scenario is related to the communication strategy in emergency situations. Usually, in emergency or disastrous situations, a small number of BSs or APs remain active, and the PSNs suffer from extreme congestion. Therefore, the heterogeneous wireless system that is based on the remaining active BSs and APs becomes an essential alternative to PSNs. Hence, in order to prioritize the data traffic of medical, security, and emergency users, those users should be assigned to different SLs according to their priority. Normally, ordinary commercial users are assigned to the lowest SL in this case.

Several other dynamic scenarios could be also considered in order to assign user priorities. For example, users could be categorized according to their MT's battery status. In order to ensure that MTs in the most critical battery status category are associated to the access technology that requests the lowest power consumption, those MTs are assigned the highest priority, and their corresponding power consumption weight (w_m^{pc}) is set to one.

Note that it is beyond the scope of this paper to discuss the advantages/disadvantages or the performance of each scenario. Instead, the problem formulated and solved in this paper could be applied in any scenario having different user priorities. Moreover, the formulated problem could be flexibly reconfigured to meet operator's objectives. For example, if it is requested to ensure that MTs with lowest SL are not always blocked upon extreme congestion, a specific number of resources in each network could be reserved for MTs with lowest SL. This can be configured in problem **P3**, and subsequently problem **P4**, by deducting in constraint (21b) the number of resources that should be reserved for MTs with lowest priority in network n. In other words, assuming that the number of resources that should be reserved in network n for MTs with lowest priority is denoted by D_n, then, "$\zeta_n - \varpi_n^{k^+}$" in constraint (21b) is replaced by "$\zeta_n - \varpi_n^{k^+} - D_n$" if $k \neq 1$ ($k = 1$ indicates the lowest SL).

5. Performance Evaluation

In Section 3.2 we have formulated a novel user association and resource allocation problem that aims at optimizing the user-centric experience in HWNs. The novelty of the formulated problem is two-folded; Firstly, it considers the data rate requirements of each user and the technology-specific resource allocation constraints of the networks. Secondly, the formulated problem considers the case where users have different priorities. The optimal solution of the formulated problem has an exponential complexity. Therefore, a solution with tolerable complexity and near-optimal performance should be proposed. Similar problems in the literature are solved following the continuous-relaxation techniques. The main contribution in the paper is the proposition of new solution with low complexity to approximate the optimal solution (Section 3.5). The relaxation-based solution discussed in Section 3.4 is implemented to study the effect of the continuous-relaxation methodology and to compare its performance to the optimal solution and to the approximation-based solution. Finally, a simple greedy solution is proposed in Section 3.5.2 to argue whether the approximation-based solution provides remarkable performance advantages when compared to a simpler solution.

In this section, we compare the performance of the different solutions discussed in Section 3 to solve, or approximate, problem **P3**. Moreover, the performance of the profit-function-based solution (Section 4.2) is evaluated. Specifically, we study, for each SL, the effect of increasing the number of active MTs on the average values of profit function, user satisfaction, signal quality, and instantaneous power consumption, in addition to the percentage of the blocked (unserved) data rate.

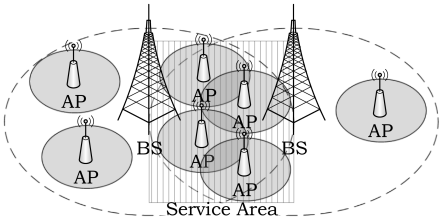

Figure 6: HWN made of LTE BSs and Wi-Fi APs. The dashed area is the service area where we focus the simulation.

The discussed solutions are evaluated through a Java language implementation and the linear program solver GLPK (for the branch and bound algorithm). A dedicated Java-based simulator is developed based on the system model presented in Section 2 to conduct the performance evaluations and comparisons.

5.1. Simulation Parameters

The simulation environment consists of two overlapping LTE BSs and four Wi-Fi APs within the service area (SA) (dashed area in Fig. 6). Focusing on the SA allows us to test all the cases without having to simulate a very large number of networks. That is, we can test the case when MTs are placed near the boundaries of the cell, and when some MTs are in a single or multiple networks coverage area. Each MT is assumed to handle only one session. The characteristics of both access technologies related to power consumption [28] and coverage range [34] are listed in Table 2.

The number of available RBs at each BS is $C_n = 75$ and the transmission power per RB is $P_n = 26$ dBm. The bandwidth of one RB is $B_n^{RB} = 180$ kHz and the noise power at all the receivers in LTE is set to -111.45 dBm [19], which corresponds to the thermal noise at room temperature and bandwidth of 180 kHz. The path loss between the LTE BS and a MT is modeled as $L(d_{mn}) = 34 + 40\log_{10}(d_{mn})$ [19]. A scheduling interval of 1 second is considered in the simulations [19] and $T_n^{BS} = 1000$ in LTE BSs. Hence, the duration of one time slot is 1 millisecond which is the duration of one transmission time interval (TTI) in the LTE standard. Concerning Wi-Fi APs, a total bandwidth $B_n = 1000$ kHz is considered for each AP, with a total transmission power of 23 dBm. The path loss model is $38.2 + 30\log_{10}(d_{mn})$ and the noise power at the MT is -90 dBm [24]. The scheduling interval is also 1 second and it is divided to 10000 time slots. MTs are randomly distributed within the SA. It is assumed that, for each MT, w_m^s and w_m^{pc} takes any random value in $[0.1, 0.9]$ such that $w_m^s + w_m^{pc} = 1$.

The simulation of each algorithm is repeated for 10^4 iterations in a Monte Carlo manner. In each iteration, the number of MTs increases from 30 to 138 one MT at a time. Thus, the system-wide optimization problem is processed 138-30=108 times in each iteration, where each system-wide decision is taken upon adding a new MT to the system. Moreover, in each iteration, the location of Wi-Fi APs and MTs changes randomly within the SA, and the initial seed of the random number generator also changes. Hence, each simulated algorithm is processed $108 * 10^4$ times, each time with different variables (number of MTs, MTs' preferences, and location of MTs and APs).

Table 2
Network Characteristics

	R_n(m)	α_n(mW/kbps)	ψ_n(mW)
LTE	500	0.05197	1288.04
Wi-Fi	200	0.13701	132.86

Table 3
Multiple Data Rates (kbps) for Different Applications

Voice call	Codec	G.729	G.726	G.711
	Datarate	32	56	87
Video call	Quality	Normal	Good	HD
	Datarate	300	500	1200
File download	Speed	Slow	Medium	Fast
	Datarate	150	700	1000

Therefore, simulated algorithms have been extensively tested in a dynamic environment in order to make sure that the collected simulation results are reliable. Hence, the general conclusions drawn out in this paper concerning the best solution of problem **P3** should not be affected by the mobility model adopted by MTs.

Each active MT randomly selects one of the data rates listed in Table 3. The service provider provides three different SLs ($K = 3$). The number of MTs subscribing to each SL is the same, *i.e.* $|\theta_1| = |\theta_2| = |\theta_3|$. Since the time required to find the optimal solution based on the branch and bound algorithm increases exponentially as the number of MTs increases, the bandwidth of APs is small to limit the number of simulated MTs.

5.2. Evaluation Metrics

The following metrics are used to evaluate the proposed solutions: average profit, average satisfaction, average signal quality, average instantaneous power consumption, and blocking percentage. The satisfaction of MT m when associated with network n is:

$$\rho_{mn} = \frac{f_{mn}}{f_{mn'}} \tag{28}$$

where n' is the index of the network for which MT m achieves the highest profit.

In fact, studying the average value of an attribute is not straightforward in a scenario where MTs request different amounts of data rate. For example, the average profit per user, *i.e.* $\frac{\sum_{m\in\theta_k}\sum_{n\in\mathcal{N}} f_{mn}x_{mn}}{|\theta_k|}$, could be increased through increasing the profit of MTs with low data rate requirements on the expense of other MTs. Thus, to avert deceptive results, the average profit per requested data rate is studied according to the following formula:

$$\frac{\sum_{m\in\theta_k}\sum_{n\in\mathcal{N}} f_{mn}x_{mn}}{\sum_{m\in\theta_k} Q_m} \tag{29}$$

Similarly, the average satisfaction per requested data rate is studied according to the following formula:

$$\frac{\sum_{m\in\theta_k}\sum_{n\in\mathcal{N}} \rho_{mn}Q_m x_{mn}}{\sum_{m\in\theta_k} Q_m} \tag{30}$$

Since ρ_{mn} represents a normalized value, it is multiplied by Q_m in the above formula.

For the signal quality and power consumption, the average values per served data rate are considered because there is no mean to calculate these values for the blocked data rates. For example, setting 0 for the power consumption of blocked data rate will decrease the average consumed power and contribute misleading results. Therefore, for power consumption, the average value per served kbps, in mW/kbps, is:

$$\frac{\sum_{m\in\theta_k}\sum_{n\in\mathcal{N}} pc_{mn}x_{mn}}{\sum_{m\in\theta_k}\sum_{n\in\mathcal{N}} Q_m x_{mn}} \tag{31}$$

Since the value of the signal quality is not related to the requested data rate, it is multiplied by Q_m to reflect the actual signal quality per served data rate. Thus the average relative received signal quality per served data rate is:

$$\frac{\sum_{m\in\theta_k}\sum_{n\in\mathcal{N}} s_{mn}Q_m x_{mn}}{\sum_{m\in\theta_k}\sum_{n\in\mathcal{N}} Q_m x_{mn}} \tag{32}$$

5.3. Simulation Results

As we have mentioned before, we will study the performance of the profit-function-based solution (Section 4.2), the optimal solution based on the branch and bound algorithm (Section 3.3), the relaxation-based solution (Algorithm 1), the approximation-based solution (Algorithm 2), and the greedy solution (Algorithm 3). It is indispensable to note that for the greedy and approximation-based solutions, the efficiency is considered as a desirability factor, *i.e.* $\Omega_{mn} = e_{mn}$, except for Section 5.3.5 where the difference between using the efficiency and normalized profit is discussed.

5.3.1. Multiple service levels

First, concerning the effect of providing different SLs, it is obvious from Fig. 7 and Fig. 8 that the proposed scheme maintains better profit and satisfaction for high-priority users. For example, as the number of MTs reaches 138, the average satisfaction is approximately 0.99, 0.6, and 0.23 for MTs with SLs 3, 2, and 1 respectively (Fig. 8). Therefore, a remarkable increase in satisfaction is maintained upon subscribing to higher SL. The same aspect is observed for the signal quality (Fig. 9) and instantaneous power consumption (Fig. 10). Moreover, Fig. 11 shows that while some MTs with SL 1 are not served, MTs with SLs 3 and 2 do not experience any blockage.

5.3.2. General behavior of algorithms

In general, increasing the number of MTs in the system strengthens the competition to acquire the limited resources of networks. Therefore, the opportunity that MTs connect to their preferred network decreases. Consequently, MTs experience degraded service illustrated by the decrease in profit and satisfaction as shown in Fig. 7 and Fig. 8 respectively. It is important to note that the average profit is 0.55 when the number of MTs is 30 because the profit function normalizes attributes through dividing them by the global maximum, *i.e.* $\max_{\forall m,n}$, which will only lead to a profit value of 1 when a single MT exists in

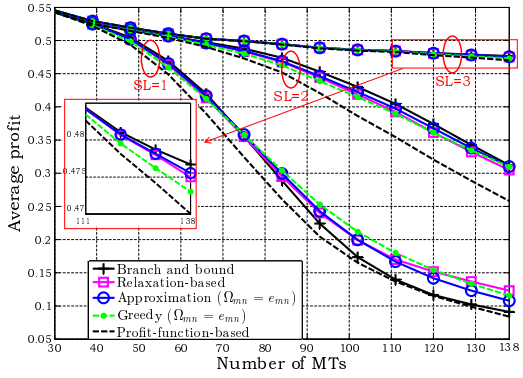

Figure 7: Average profit per requested data rate (according to Eq. (29)).

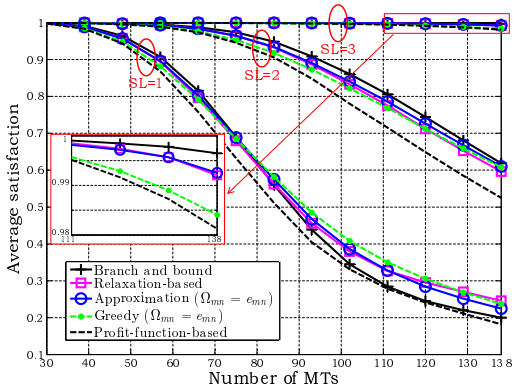

Figure 8: Average satisfaction per requested data rate (according to Eq. (30)).

the system. However, this behavior does not impact the user satisfaction. In order to increase the overall profit, the optimization problem **P3** finds the best set of association values for all MTs. Thus, the main performance of the optimization problem and its different solutions could be studied through the profit, and consequently through the satisfaction because it is directly related to the profit. It is shown in Fig. 7 and Fig. 8 that the proposed approximation-based solution and the relaxation-based solution maintain performance near the optimal solution for MTs with SL 3. The greedy solution, although it tends to approach the optimal solution, performs near the profit-function-based solution which has the worst performance. Therefore, the proposed approximation-based solution efficiently approximates the optimal solution, and could overwhelm the relaxation-based solution. Concerning MTs with SL 2, as the number of MTs increases, the relaxation-based, approximation-based, and greedy solutions perform near the optimal solution, and far away from the profit-function-based solution. It is remarkable that the proposed approximation-based solution maintains the nearest performance to the optimal one (Fig. 7 and Fig. 8).

Optimal resource allocation for high-priority MTs causes efficient resource utilization in networks. Hence, the chance that

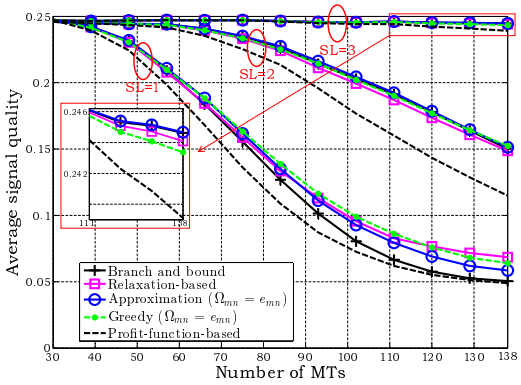

Figure 9: Average relative recieved signal strength per served data rate (according to Eq. (32)).

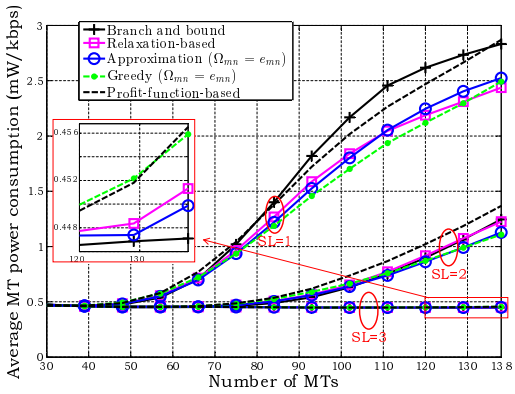

Figure 10: Average power consumption per served data rate (according to Eq. (31)).

MTs with low priority associate to their preferred network decreases. For example, Wi-Fi APs are usually preferred for their low power consumption feature; efficient resource utilization in these APs lowers the number of unallocated resources, which in turns lowers the chance that MTs with low priority associate to these APs. Consequently, upon adopting the optimal solution, MTs with SL 1 experience service near the profit-function-based solution. This is recognized in Fig. 7 and Fig. 8 where the optimal solution starts approaching the profit-function-based solution as the number of MTs increases beyond 80.

As a matter of fact, the profit function depends on the location of the MT, the requested data rate, the normalized values of the signal quality and power consumption, and the weights w_m^s and w_m^{pc} which are different between MTs. Therefore, it is normal not to notice the same behavior of the satisfaction curve (Fig. 8) reflected in the curves of the signal quality and power consumption (Fig. 9 and Fig. 10 respectively). However, the satisfaction could reflect a general behavior of the compared solutions in terms of signal quality and power consumption. For example, Fig.9 and Fig. 10 show that the proposed approximation-based solution maintains near-optimal performance for MTs with SL 3. This is illustrated through high signal quality and low instantaneous power consumption. Moreover, the degraded service of the optimal solution for MTs with SL 1 in Fig.9 and Fig. 10 is a result for the same reason discussed before for the profit and satisfaction of those MTs.

Considering the difference between the highest and the second highest available desirability value (*line 2.22* in Algorithm 2), and maintaining a semi-global view on the system explain the significant results contributed by the proposed approximation-based solution.

5.3.3. Blocking percentage evaluation

In order to fully understand the behavior of the proposed solutions, the percentage of blocked data rate should be studied. According to the proposed solution, the HOE has the privilege to reassign resources used by low-priority MTs to MTs with higher priorities. Therefore, MTs with SLs 3 and 2 do not suffer from any blockage throughout the simulation. Concerning MTs with SL 1, Fig. 11 illustrates that the optimal solution achieves the lowest data rate blockage. The proposed approximation-based solution maintains lower blocking percentage than the relaxation-based, greedy and profit-function-based solutions. For instance, the profit-function-based solution suffers from 16% blockage when the number of MTs reaches 138. The greedy solution lowers down this percentage to 9, followed by the relaxation-based and approximation-based solutions that score 8.1% and 7.8% respectively, while the optimal solution scores about 7.2%. Therefore, as can be seen in Fig. 11, the proposed approximation-based solution achieves and maintains the lowest blocking percentage among the tested approaches, except for the optimal one of course. Such result is considered as a major improvement since users subscribing to the lowest SL would be mainly concerned about having a service, without paying much attention to the performance.

Since the efficiency e_{mn} factor accounts for the channel conditions, then considering e_{mn} as a desirability factor plays a vital role in decreasing the data rate blocking percentage for both the greedy and the approximation-based solutions. Moreover, the proposed approximation-based solution achieves low blocking percentage for prioritizing those MTs with a single available network, *i.e.* $|F_m| = 1$ in Algorithm 2, among other MTs.

5.3.4. Complexity-performance trade off

We are interested in comparing the performance of our proposed approximation-based solution to the optimal performance that could be achieved upon following the continuous relaxation approach, *i.e.* Algorithm 1. In addition, we are interested in studying the impact of the continuous relaxation on the performance of the algorithm.

The complexity of Algorithm 1, *i.e.* the relaxation-based solution, could not be determined because it is impossible to analytically determine the number of times the linear program will be solved while emptying $\mathcal{S}_2^k$. Note that a relaxation-based solution with determined complexity could be proposed by simply solving problem **P4** once, and converting the fractional association values into boolean using some heuristic. However, the

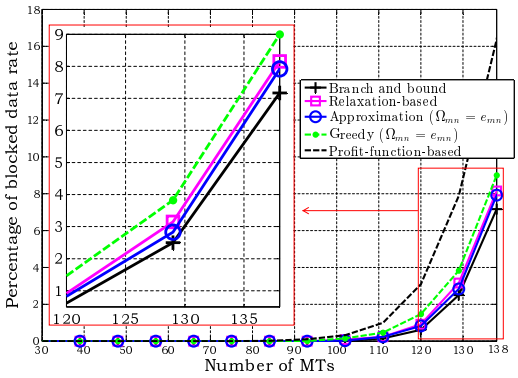

Figure 11: Percentage of the blocked data rate for MTs with SL 1.

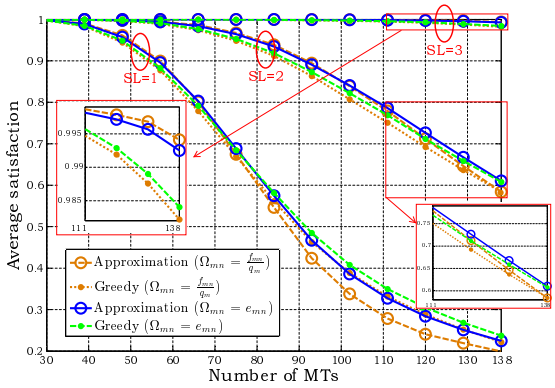

Figure 12: Average satisfaction per requested data rate (according to Eq. (29)). Comparing the difference between using the efficiency and normalized profit as desirability factor.

Table 4
Solution Complexity

	Complexity
P3	$O(N^{\lvert\theta_k\rvert})$
Solving **P4**	$O(\lvert\theta_k\rvert^3 N^3 L)$
Algorithm 2	$O(\lvert\theta_k\rvert^2 + \lvert\theta_k\rvert N \log N)$
Algorithm 3	$O(\lvert\theta_k\rvert N + \lvert\theta_k\rvert N \log \lvert\theta_k\rvert N)$

performance of such solution will not be better than that of Algorithm 1 which is considered as the optimal solution based on the continuous relaxation methodology. Moreover, the complexity of such solution will be higher than that of solving the relaxed problem **P4**, *i.e.* $O(|\theta_k|^3 N^3 L)$, due to the additional heuristic step.

Anyways, the complexity of the relaxation-based solution could be expressed in terms of the cube of the number of MTs in each SL multiplied to the cube of the number of networks. Where as, the complexity of the optimal solution based on the branch and bound algorithm is the highest among all the proposed solutions because it is exponential.

On the other hand, the complexity of the approximation-based solution (Algorithm 2) is mainly related to the square of the number of MTs in each SL. So, it is obvious that the complexity of Algorithm 2 is less than that of solving the relaxed problem **P4**. The greedy solution has the lowest complexity because it is mainly related to the number of MTs in each SL multiplied to the number of networks. Table 4 lists all the discussed complexities sorted according to their decreasing order.

Simulation results discussed earlier show that relaxing the binary constraint causes degradation in the performance when compared to the optimal solution. Even the approximation-based solution, which has a complexity lower than that of solving the relaxed problem **P4**, could perform similar (Fig. 7), and sometimes better (Fig. 11) than the relaxation-based solution. Although the approximation-based solution requires higher complexity than the greedy one, but the former has shown more robustness mainly in terms of the data rate blockage and the performance of MTs with SL 3. Therefore, the approximation-based solution demonstrates a remarkable trade

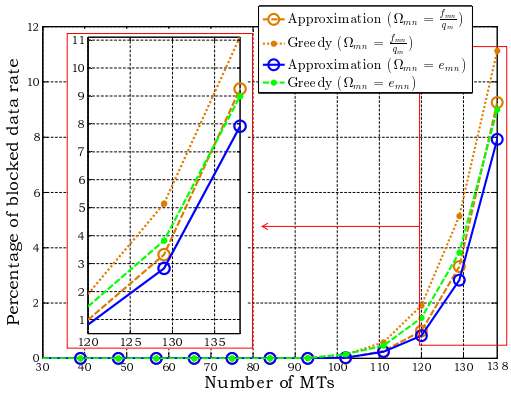

Figure 13: Percentage of the blocked data rate for MTs with SL 1. Comparing the difference between using the efficiency and normalized profit as desirability factor.

off between the complexity on one side and the performance on the other side.

5.3.5. *Efficiency factor verses normalized profit*

The difference in satisfaction between using the normalized profit, *i.e.* $\frac{f_{mn}}{Q_m}$, and the efficiency in the approximation-based and greedy solutions is shown in Fig. 12. In fact, associating a MT to the network with the highest efficiency does not guarantee the highest profit. Instead, it guarantees the highest profit per single allocated bandwidth unit, *i.e.* 1 Hz. Therefore, it is normal to notice in Fig. 12 that using the normalized profit instead of the efficiency in the approximation-based solution contributes higher profit for MTs with SL 3. Actually, the effect of using the efficiency is enlarged because it is considered twice in Algorithm 2 where the difference between the highest and the second highest efficiency is used to take a decision. However, as the number of MTs increases, adopting the efficiency as desirability factor contributes higher profit as shown in Fig. 12 for SLs 2 and 1 because accounting for the requested

bandwidth upon congestion is essential. Moreover, using the efficiency lowers the blocking percentage significantly for both the greedy and the approximation-based solutions as shown in Fig. 13. Hence, the efficiency is chosen as the main desirability measure.

6. Conclusion

In this paper, we have proposed a solution for the priority-based user association and downlink resource allocation problem in a heterogeneous wireless system. The proposed solution considers the user preferences and the data rate requested by each user. First, we have formulated an optimization problem and then simplified it. The formulated problem prevents MTs with low priority from utilizing resources of high-priority MTs. For the simplified problem, we discussed the relaxation-based and greedy solutions, and proposed a novel approximation-based solution. We also proposed a solution management strategy to reduce the number of SLs that the optimization function will process. Simulation results encourage users to subscribe to the highest priority where they experience the best service. Concerning the proposed solutions, simulation results show that the proposed approximation-based solution maintains performance near the optimal one. Therefore, operators are encouraged to adopt the proposed approximation-based solution to maintain better service for users, and to increase their economical profit through reducing the data rate blockage.

This paper sheds the light on the importance of exploring user demands and preferences within HWNs. Moreover, the formulated problem and proposed solutions pave the way for an optimized ABC scheme.

References

References

[1] R. Q. Hu, Y. Qian, An energy efficient and spectrum efficient wireless heterogeneous network framework for 5g systems, IEEE Communications Magazine 52 (5) (2014) 94–101. doi:10.1109/MCOM.2014.6815898.

[2] L. Wang, G. S. G. S. Kuo, Mathematical modeling for network selection in heterogeneous wireless networks-a tutorial, IEEE Communications Surveys Tutorials 15 (1) (2013) 271–292. doi:10.1109/SURV.2012.010912.00044.

[3] E. Gustafsson, A. Jonsson, Always best connected, IEEE Wireless Communications 10 (1) (2003) 49–55. doi:10.1109/MWC.2003.1182111.

[4] J. B. Ernst, S. C. Kremer, J. J. P. C. Rodrigues, Heterogeneous wireless network rat selection with multiple operators and service contracts, in: IEEE International Conference on Communications (ICC), 2015, pp. 6011–6017. doi:10.1109/ICC.2015.7249280.

[5] D. Xenakis, N. Passas, L. Merakos, C. Verikoukis, Archon: An andsf-assisted energy-efficient vertical handover decision algorithm for the heterogeneous ieee 802.11/lte-advanced network, in: IEEE International Conference on Communications (ICC), 2014, pp. 3166–3171. doi:10.1109/ICC.2014.6883808.

[6] A. Hasswa, N. Nasser, H. Hassanein, Tramcar: A context-aware cross-layer architecture for next generation heterogeneous wireless networks, in: IEEE International Conference on Communications (ICC), Vol. 1, 2006, pp. 240–245. doi:10.1109/ICC.2006.254734.

[7] A. Awad, A. Mohamed, C. F. Chiasserini, User-centric network selection in multi-rat systems, in: IEEE Wireless Communications and Networking Conference Workshops (WCNCW), 2016, pp. 97–102. doi:10.1109/WCNCW.2016.7552682.

[8] R. Trestian, O. Ormond, G. M. Muntean, Energy-quality-cost tradeoff in a multimedia-based heterogeneous wireless network environment, IEEE Transactions on Broadcasting 59 (2) (2013) 340–357. doi:10.1109/TBC.2013.2244790.

[9] B. H. Jung, N. O. Song, D. K. Sung, A network-assisted user-centric wifi-offloading model for maximizing per-user throughput in a heterogeneous network, IEEE Transactions on Vehicular Technology 63 (4) (2014) 1940–1945. doi:10.1109/TVT.2013.2286622.

[10] Q. Ye, B. Rong, Y. Chen, C. Caramanis, J. G. Andrews, Towards an optimal user association in heterogeneous cellular networks, in: IEEE Global Communications Conference (GLOBECOM), 2012, pp. 4143–4147. doi:10.1109/GLOCOM.2012.6503766.

[11] T. Zhou, Y. Huang, L. Yang, User association with jointly maximising downlink sum rate and minimising uplink sum power for heterogeneous cellular networks, IET Communications 9 (2) (2015) 300–308. doi:10.1049/iet-com.2014.0476.

[12] C. Liu, P. Whiting, S. V. Hanly, Joint resource allocation and user association in downlink three-tier heterogeneous networks, in: IEEE Global Communications Conference (GLOBECOM), 2014, pp. 4232–4238. doi:10.1109/GLOCOM.2014.7037472.

[13] S. Borst, S. Hanly, P. Whiting, Optimal resource allocation in hetnets, in: IEEE International Conference on Communications (ICC), 2013, pp. 5437–5441. doi:10.1109/ICC.2013.6655454.

[14] Z. Du, Q. Wu, P. Yang, Y. Xu, J. Wang, Y. D. Yao, Exploiting user demand diversity in heterogeneous wireless networks, IEEE Transactions on Wireless Communications 14 (8) (2015) 4142–4155. doi:10.1109/TWC.2015.2417155.

[15] L. Chen, H. Li, An mdp-based vertical handoff decision algorithm for heterogeneous wireless networks, in: IEEE Wireless Communications and Networking Conference, 2016, pp. 1–6. doi:10.1109/WCNC.2016.7564804.

[16] Y. Chen, J. Li, Z. Lin, G. Mao, B. Vucetic, User association with unequal user priorities in heterogeneous cellular networks, IEEE Transactions on Vehicular Technology 65 (9) (2016) 7374–7388. doi:10.1109/TVT.2015.2488039.

[17] C. Tata, M. Kadoch, Efficient priority access to the shared commercial radio with offloading for public safety in lte heterogeneous networks, Journal of Computer Networks and Communications.

[18] X. Yang, J. Bigham, L. Cuthbert, Resource management for service providers in heterogeneous wireless networks, in: IEEE Wireless Communications and Networking Conference, 2005, Vol. 3, 2005, pp. 1305–1310 Vol. 3. doi:10.1109/WCNC.2005.1424705.

[19] H. Boostanimehr, V. K. Bhargava, Unified and distributed qos-driven cell association algorithms in heterogeneous networks, IEEE Transactions on Wireless Communications 14 (3) (2015) 1650–1662. doi:10.1109/TWC.2014.2371465.

[20] T. S. Rappaport, Wireless communications: principles and practice, Vol. 2, Prentice Hall PTR New Jersey, 1996.

[21] Q. Han, B. Yang, G. Miao, C. Chen, X. Wang, X. Guan, Backhaul-aware user association and resource allocation for energy-constrained hetnets, IEEE Transactions on Vehicular Technology 66 (1) (2017) 580–593. doi:10.1109/TVT.2016.2533559.

[22] B. G. Lee, D. Park, H. Seo, Wireless communications resource management, John Wiley & Sons, 2009.

[23] P. Y. Kong, G. K. Karagiannidis, Backhaul-aware joint traffic offloading and time fraction allocation for 5g hetnets, IEEE Transactions on Vehicular Technology 65 (11) (2016) 9224–9235. doi:10.1109/TVT.2016.2517671.

[24] P. Xue, P. Gong, J. H. Park, D. Park, D. K. Kim, Radio resource management with proportional rate constraint in the heterogeneous networks, IEEE Transactions on Wireless Communications 11 (3) (2012) 1066–1075. doi:10.1109/TWC.2011.102611.110281.

[25] J. Choi, J. Yoo, S. Choi, C. Kim, Eba: an enhancement of the ieee 802.11 dcf via distributed reservation, IEEE Transactions on Mobile Computing 4 (4) (2005) 378–390. doi:10.1109/TMC.2005.57.

[26] S. Kim, S. Choi, B. G. Lee, A joint algorithm for base station operation and user association in heterogeneous networks, IEEE Communications Letters 17 (8) (2013) 1552–1555. doi:10.1109/LCOMM.2013.070113.130730.

[27] W. Shen, Q. A. Zeng, Cost-function-based network selection strategy in integrated wireless and mobile networks, IEEE Transactions on Vehicular

Technology 57 (6) (2008) 3778–3788. doi:10.1109/TVT.2008.917257.

[28] J. Huang, F. Qian, A. Gerber, Z. M. Mao, S. Sen, O. Spatscheck, A close examination of performance and power characteristics of 4g lte networks, in: Proceedings of the 10th international conference on Mobile systems, applications, and services, ACM, 2012, pp. 225–238.

[29] S. Sadr, R. S. Adve, Partially-distributed resource allocation in small-cell networks, IEEE Transactions on Wireless Communications 13 (12) (2014) 6851–6862. doi:10.1109/TWC.2014.2327030.

[30] R. M. Karp, Reducibility among combinatorial problems, in: Complexity of computer computations, Springer, 1972, pp. 85–103.

[31] F. A. Potra, S. J. Wright, Interior-point methods, Journal of Computational and Applied Mathematics 124 (1) (2000) 281–302.

[32] S. Martello, P. Toth, An algorithm for the generalized assignment problem, Operational research 81 (1981) 589–603.

[33] Ieee standard for local and metropolitan area networks - media independent handover services, IEEE Std 802.21-2008doi:10.1109/IEEESTD.2009.4769367.

[34] Q. T. Nguyen-Vuong, N. Agoulmine, E. H. Cherkaoui, L. Toni, Multicriteria optimization of access selection to improve the quality of experience in heterogeneous wireless access networks, IEEE Transactions on Vehicular Technology 62 (4) (2013) 1785–1800. doi:10.1109/TVT.2012.2234772.