Licensed and Unlicensed Spectrum Allocation in Heterogeneous Networks

Zhiyi Zhou, Dongning Guo, and Michael L. Honig
Department of Electrical Engineering and Computer Science
Northwestern University, Evanston, IL 60208, USA

Abstract

In future networks, an operator may employ a wide range of access points using diverse radio access technologies (RATs) over multiple licensed and unlicensed frequency bands. This paper studies centralized user association and spectrum allocation across many access points in such a heterogeneous network (HetNet). For practical reasons, such centralized control is likely to be on a much slower timescale than frame transmissions. A queueing model is introduced to capture the lower spectral efficiency, reliability, and additional delays of data transmission over the unlicensed bands due to contention and/or listen-before-talk requirements. Two optimization-based spectrum allocation schemes are proposed along with efficient algorithms for computing the allocations. The proposed solutions are fully aware of traffic loads, network topology, as well as external interference levels in the unlicensed bands. Packet-level simulation results show that the proposed schemes significantly outperform orthogonal and full-frequency-reuse allocations under all traffic conditions.

Index Terms

LTE, LTE in unlicensed spectrum (LTE-U), multiple radio access technologies, spectrum allocation, user association, utilization.

I. INTRODUCTION

The wireless industry has set an ambitious goal to increase the area capacity (in bits per second per square meter) by three orders of magnitude in the next five to ten years. In addition

This work was supported in part by a gift from Futurewei Technologies and by the National Science Foundation under Grant No. CCF-1423040.
to densely deploying small cells and improving the spectral efficiency [2], another avenue is to exploit all available spectrum, which is a relatively scarce resource [3]. Future generation cellular networks are likely to involve multiple radio access technologies (RATs) over multiple frequency bands. Such RATs include Long Term Evolution (LTE), WiFi, and LTE in unlicensed spectrum (LTE-U). The prime bands today are licensed frequency bands under 3 GHz and over 500 MHz of unlicensed spectrum in the 2.4 GHz and 5 GHz frequency bands.

It is important to allocate all physical resources in a spectrally efficient and computationally efficient manner. Current 4G cellular networks generally use regular frequency reuse patterns, including full frequency reuse and fractional frequency reuse. In the former scheme, every cell uses all available frequency bands, whereas the latter scheme is similar, except that cells use orthogonal frequency bands at the cell edge to reduce mutual interference. However, such simple methods are not as effective in emerging heterogeneous networks (HetNets), whose topologies are often highly irregular, and whose small cells have highly dynamic traffic conditions.

There have been many studies on spectrum allocation in cellular networks (see, e.g., [4]–[11]). Most authors formulate the allocation problem as that of deciding, for each slice of the spectrum, which access points (APs) and/or links should use it. In addition, user association was considered in [7], [8]. In general, hard nonconvex integer programming problems are formulated, which typically have many local optima. Furthermore, the figures of merit used in most work are physical layer performance measures such as sum rate and outage probability. These traffic-independent metrics may not reflect a user’s true quality of service (QoS) in HetNets with large traffic variations in overlapping cells with complicated interference conditions.

To address the preceding issues, an alternative optimization-based framework was developed in [12] and [13] for allocating downlink spectrum resources in a HetNet. This framework allows arbitrary user association and flexible spectrum allocation, driven entirely by users’ traffic loads. This framework is well matched to emerging joint processing architectures such as cloud radio access network (C-RAN) and cell-free massive multiple-input multiple-output systems [14]. Here, the timescale of resource adaptation is conceived to be once every a few seconds or minutes in the context of current networks. This timescale is, on the one hand, fast enough for tracking the traffic variation, and, on the other hand, slow enough to allow joint optimization of many APs with a large number of user equipments (UEs). Another advantage is that the spectrum resources can be assumed to be homogeneous on this timescale, namely, every hertz can support
the same spectral efficiency when averaged over many frames. The approach in [12] has been generalized to incorporate energy-efficient allocation via cell activation [13] as well as the effect of opportunistic scheduling on a fast timescale [15], [16].

In this work we generalize the framework of [12] and [13] to the scenario in which there are multiple RATs over multiple frequency bands. Different queueing models are proposed to distinguish the characteristics of licensed and unlicensed bands. The multiple-band allocation problem is formulated as a bi-convex optimization problem. A conservative allocation scheme is first proposed following the approach of [12] and [13]. This is followed by a utilization-dependent allocation scheme which incorporates the utilizations of the APs into the formulation to more accurately account for dynamic inter-cell interference. An iterative algorithm is designed to solve the allocation problem with manageable computational complexity for small systems. In each step, we solve a convex problem with a unique optimal solution.

Another feature of this work is that the packet length is allowed to have a general probability distribution. In prior works [12], [13], [15] and [16], the packet length is assumed to be exponentially distributed to yield a simple analytic form in the objective representing the QoS. We show that the proposed formulation and algorithm apply to general packet length distributions and hence a broad class of traffic conditions.

To evaluate the performance, packet-level simulations are carried out. It is shown that both the conservative allocation scheme and the utilization-based allocation scheme significantly reduce the average packet delay in the heavy traffic regime compared to orthogonal allocation and full-frequency-reuse allocation. The large performance gain is observed mainly because the proposed schemes are traffic-aware and also exploit the particular characteristics of each RAT. In addition, the utilization-dependent allocation scheme attains the best performance over all traffic conditions due to its accurate modeling of the dynamic interference among APs.

The remainder of this paper is organized as follows. The system model is introduced in Section II. The optimization problem using the conservative allocation is formulated in Section III. The utilization-dependent allocation scheme is presented in Section IV. The extension to general packet length distributions is given in Section V. Simulation results are presented in Section VI and conclusions are presented in Section VII. All technical proofs are relegated to the appendices.
II. System Model

In this section, we introduce models for user traffic, spectrum allocation, and link throughput in a similar manner as in [13]. A new queueing model is also introduced for traffic over unlicensed bands.

A. Spectrum Allocation

We consider the downlink of a heterogeneous network consisting of \( n \) APs. Without loss of generality, suppose each AP can use all \( m \) different radio access technologies, with each RAT on a separate frequency band. To reduce complexity, the UEs are divided into \( k \) groups based on their location, average channel condition, and delay requirements (if applicable), so that UEs in each group have similar channel conditions on the slow timescale. This is reasonable because it suffices to carry out the slow timescale allocation to the UE groups rather than individual UEs. It is understood that fast timescale physical resource allocation is carried out by schedulers (usually one scheduler in each AP).

Denote the set of all APs by \( \mathcal{N} = \{1, ..., n\} \), the set of all user groups by \( \mathcal{K} = \{1, ..., k\} \), and the set of RATs by \( \mathcal{M} = \{1, ..., m\} \). RAT \( l \) employs its separate homogeneous spectrum of bandwidth \( W_l \). We allow arbitrary association so that each AP can simultaneously serve any subset of UE groups and each UE can be simultaneously served by any subset of APs. Furthermore, we allow flexible resource allocation in that each AP-UE link can use an arbitrary number of RATs, where each RAT uses an arbitrary (possibly discontinuous) subset of the available spectrum. Despite the enormous number of possibilities, we will show that the actual AP-UE association and spectrum allocation is extremely sparse.

The key to total spectrum agility is the notion of pattern [12], [13]. In general, a pattern simply refers to a subset of transmitters. A time-frequency resource is said to be reserved for pattern \( \mathcal{A} \) if the resource is to be shared by transmitters in \( \mathcal{A}[1] \). In the downlink, a pattern \( \mathcal{A} \) is a subset of \( \mathcal{N} \), and all APs in \( \mathcal{A} \) have simultaneous access to the frequency bands associated with pattern \( \mathcal{A} \). Assuming known transmit power spectral densities (PSDs), the pattern of a resource determines the signal-to-interference-plus-noise ratio (SINR) and hence the spectral efficiencies

\[ \text{SINR} = \frac{P_t}{I + N_0} \]

\[ P_t \] is the transmit power, \( I \) is the interference, and \( N_0 \) is the noise power. The notion of pattern finds its root in the concept of independent set defined in the special case where the network is described by a conflict graph, and where nodes/links in the independent set share the same resources [17].
of all links over the resource. The allocation problem can then be formulated as how to divide the spectrum of each RAT among all \(2^n - 1\) nonempty patterns. To illustrate the concept of pattern with an example, suppose there are two APs operating over one frequency band. The spectrum can be divided into \(2^2 - 1 = 3\) segments, where one segment is used by AP 1 exclusively (the pattern is \(\{1\}\)), a second is used by AP 2 exclusively (the pattern is \(\{2\}\)), and the remaining segment is shared by the two APs (the pattern is \(\{1, 2\}\)). If the transmit pattern of a certain spectrum resource is \(\{1, 2\}\), i.e., both APs transmit, then the spectral efficiencies of all links are determined.

A network controller collects traffic load and channel/interference information from all the APs. The task of the central controller is to determine which spectrum segments to allocate to each AP, and subsequently, which sub-segments to allocate to specific UEs associated with that AP. The problem is equivalent to the combination of the following three subproblems:

1) Decide which RAT or RATs should be used to serve each UE group.

2) Allocate the bandwidth across all \(2^n\) patterns for each RAT. This is denoted by a \(2^n \times m\)-tuple: \(y = (y_{l,A})_{A \subset N, l \in M}\), where \(y_{l,A} \in [0, 1]\) is the fraction of spectrum assigned to RAT \(l\) shared by APs in set \(A\). Clearly,

\[
\sum_{A \subset N} y_{l,A} = 1, \quad \forall l \in M
\]  

(1)

and any efficient allocation would not use the empty pattern, so that \(y_{l,\varnothing} = 0, \forall l \in M\).

3) For every pattern \(A \subset N\), for every AP \(i \in A\), divide the bandwidth \(y_{l,A}\) to the UE groups. Denote the bandwidth allocated to the link \(i \to j\) over pattern \(A\) under RAT \(l\) as \(x_{l,A}^{i,j}\). Then we have:

\[
\sum_{j \in K} x_{l,A}^{i,j} \leq y_{l,A}, \quad \forall i \in A, A \subset N, l \in M.
\]  

(2)

The user association is indirectly decided by the amount of spectrum resources assigned by each AP to each group, namely, UE \(j\) is assigned to AP \(i\) over RAT \(l\) if and only if \(x_{l,A}^{i,j} > 0\) for some pattern \(A\).

B. From Spectrum to Transmission Rates

An important measure of user QoS is the average packet delay in the system. The delay is determined by the packet arrival rates and the service rates. The instantaneous service rate
of a UE group’s queue in turn depends on the spectral efficiencies along with the bandwidths allocated to that specific UE group. For simplicity, it is assumed that when AP $i$ transmits over RAT $l$, it employs all patterns available to it and applies flat PSD $p_i^l$ over the allocated spectrum. At any frequency designated for RAT $l$, the instantaneous spectral efficiency achievable by the link from AP $i$ to UE $j$ depends on the set of active APs $\mathcal{A} \subset \mathcal{N}$ using that frequency. Let this (spectral) efficiency be denoted by $s_{i \rightarrow j}^{l,\mathcal{A}}$. Clearly, $s_{i \rightarrow j}^{l,\mathcal{A}} = 0$ if $i \notin \mathcal{A}$. Moreover, the spectral efficiency decreases as more APs become active, i.e., $s_{i \rightarrow j}^{l,\mathcal{A}} \geq s_{i \rightarrow j}^{l,\mathcal{B}}$ if $i \in \mathcal{A} \subset \mathcal{B}$. On the slow timescale, the spectral efficiencies are either known a priori or can be measured and sent to the central controller. For later convenience, we normalize the spectral efficiency over RAT $l$ using factor $L/W$ (bits/second/Hz) so that the unit of $s_{i \rightarrow j}^{l,\mathcal{A}}$ is packets/second.

For concreteness in obtaining numerical results, we use Shannon’s formula to arrive at the expression for link efficiencies:

$$s_{i \rightarrow j}^{l,\mathcal{A}} = \frac{\alpha_l W_l \mathbb{1}(i \in \mathcal{A})}{L} \log_2 \left( 1 + \frac{p_i^l}{I_{i \rightarrow j}^l} \right) \text{ packets/s},$$

(3)

where $\alpha_l \in (0, 1]$ is the discount factor for RAT $l$, $\mathbb{1}(i \in \mathcal{A}) = 1$ if $i \in \mathcal{A}$ and $\mathbb{1}(i \in \mathcal{A}) = 0$ otherwise, and $I_{i \rightarrow j}^l$ is the total noise plus interference PSD from other APs in $\mathcal{A}$ to UE $j$, which depends on their transmit PSDs and path loss. The discount factor of a licensed band is typically closer to 1 than that of an unlicensed band due to external interference from other operators in the unlicensed band. The average effect of small-scale fading can be included by considering the ergodic capacity in lieu of (3), which does not change the main developments of this paper.

Over a resource reserved for pattern $\mathcal{A}$, only the subset of APs in $\mathcal{A}$ with data to transmit will be using the resource at any given time. In general, the instantaneous service rate depends on the set of transmitting APs and the rate adaptation scheme. Two service rate models are adopted as described in the following.

Under the conservative allocation introduced in [12], if some of RAT $l$’s spectrum resources reserved for pattern $\mathcal{A}$ are allocated to link $i \rightarrow j$, then AP $i$ transmits UE $j$’s packets at spectral efficiency $s_{i \rightarrow j}^{l,\mathcal{A}}$ using RAT $l$ over pattern $\mathcal{A}$. This rate is achievable even if all APs in $\mathcal{A}$ have data to transmit (hence the name conservative). This is equivalent to assuming that other APs’ traffic is always backlogged. The rate contributed by spectrum reserved for pattern $\mathcal{A}$ under RAT $l$ is the product of the spectral efficiency and the bandwidth: $s_{i \rightarrow j}^{l,\mathcal{A}} x_{i \rightarrow j}^{l,\mathcal{A}}$. The total service rate
for the queue of UE $j$ over RAT $l$ is the sum rate of all APs and patterns, expressed as:

$$r_j^l = \sum_{A \subset N} \sum_{i \in N} s_{i \rightarrow j}^{i \rightarrow j} x_{i \rightarrow j}^{i \rightarrow j} \text{ packets/s.} \quad (4)$$

The advantage of the conservative allocation is that there is no need for the scheduler (or whichever unit that performs rate adaptation) to know the state of the other APs included in the pattern. However, the service rate under conservative allocation is a lower bound on what is achievable because the interference is overestimated.

To adapt the allocation to the actual interference pattern, we introduce a new model in this work, referred to as the utilization-dependent model. Under this model, the service rate contributed by the spectrum reserved for pattern $A$ depends on the subset of transmitting APs in $A$. Specifically, the service rate for UE group $j$ over RAT $l$ when the set of active interfering APs is $I$ is expressed as:

$$r_{j,l,I} = \sum_{A \subset N} \sum_{i \in N} s_{i \rightarrow j}^{i \rightarrow j} x_{i \rightarrow j}^{i \rightarrow j} \text{ packets/s.} \quad (5)$$

C. Queueing Model

Without loss of generality, we restrict the subsequent investigation to the scenario with 2 RATs, where one RAT is over the licensed band (conceived as LTE) and the other is over the unlicensed band (conceived as LTE-U). This can be easily generalized to more than two RATs.

The traffic to UE group $j$ is modeled as an independent Poisson point process with rate $\lambda^j$ packets per second. The packet lengths are independent random variables with $L$ bits on average. Packets intended for each UE group are transmitted according to the first-in-first-out (FIFO) discipline. The buffer in each queue is assumed to be unlimited for simplicity. The traffic load of UE group $j$ is further divided into two streams served by the two RATs, respectively. This may be implemented by dividing the group of UEs for association with different RATs, so that each UE is only served by one RAT. The resulting packet arrival rates of two streams are denoted as $\lambda_1^j$ and $\lambda_2^j$, respectively, which are variables to be optimized subject to the constraint:

$$\lambda_1^j + \lambda_2^j \geq \lambda^j. \quad (6)$$

For each RAT, the corresponding $k$ UE groups form a system of $k$ interactive queues where the instantaneous service rate of each queue may depend on which other queues are empty.
The physical and/or medium access control layers of LTE-U are designed to facilitate coexistence with other RATs in unlicensed bands, such as WiFi. In [19], a listen-before-talk scheme was proposed in which carrier sensing is embedded in a deterministic portion of a LTE subframe. In [20], the authors proposed the use of LTE uplink power control to improve LTE/WiFi coexistence. References [21] and [22] propose to enable LTE/WiFi coexistence by muting LTE transmission on certain subframes following a pre-determined pattern.

In this paper, we assume that LTE-U has a listen-before-talk feature, which is likely to be a dominant mode in emerging LTE-U standards [18]. In this mode, an AP operating LTE-U with data to send first performs carrier sensing before its transmission. We use queues with vacation and nonexhaustive service [23] to model such a mechanism. The queueing scheme is depicted in Fig. 1. Specifically, the server of queue $j$ will take a single vacation after completing the service of each packet. The vacation duration $V_j$ (in seconds) is a random variable following a certain distribution. The mean and variance of the vacation time depend on the level of local interference, such as interference from WiFi users. The higher the interference level, the more time a queue has to wait before being served.

**III. QUEUEING DELAYS AND A CONSERVATIVE ALLOCATION SCHEME**

In this section, we adopt the conservative service rate model (4) to develop a scheme for allocating licensed and unlicensed spectrum. For each RAT, the corresponding $k$ UE groups form $k$ interactive queues. Since the service rate of each queue is independent of the states of other queues under the conservative service rate model (4), the $k$ interactive queues become $k$ independent M/M/1 queues. Let RAT 1 represent LTE and RAT 2 represent LTE-U.

$^2$Challenges of such coexistence are discussed in [18].
Under LTE, the average packet delay (in seconds/packet) of UE group $j$ is given by the average delay for the M/M/1 queue:

$$t_1^j = \frac{1}{(r_1^j - \lambda_1^j)^+},$$

where $(x)^+ = x$ if $x > 0$ and $(x)^+ = 0$ if $x \leq 0$. If $r_1^j \leq \lambda_1^j$, the queueing delay is infinite, i.e., the queue becomes unstable. It is important to note that the right hand side of (7) is convex in the pair $(r_1^j, \lambda_1^j)$ on $\mathbb{R}^2$.

Under LTE-U, the $k$ UE groups form $k$ independent M/M/1 queues with single vacation. We have:

**Proposition 1**: The average packet delay (in seconds/packet) of a queue with Poisson arrival of rate $\lambda$ and a single server of rate $r$ with single vacation $V$ whose expected squared duration is $\nu = E[V^2]$ is given by:

$$t = \frac{2 + r\lambda \nu}{2(r - \lambda)^+}.$$  \hspace{1cm} (8)

For fixed $\nu$, the function defined by (8) is convex in the pair $(r, \lambda)$ on $\mathbb{R}^2$.

The proof of Proposition 1 is given in Appendix A.

The joint user association and conservative spectrum allocation problem is formulated as P1:

$$\text{minimize} \quad U = \sum_{j \in K} \lambda_j^1 t_1^j + \lambda_j^2 t_2^j$$

subject to

$$t_1^j = \frac{1}{(r_1^j - \lambda_1^j)^+}, \quad j \in K \hspace{1cm} (P1b)$$

$$t_2^j = \frac{2 + r_2^j \lambda_2^j \nu^j}{2(r_2^j - \lambda_2^j)^+}, \quad j \in K \hspace{1cm} (P1c)$$

$$r_l^j = \sum_{A \subset N} \sum_{i \in N} x_{i \rightarrow j, A}^l, \quad j \in K; l \in \{1, 2\} \hspace{1cm} (P1d)$$

$$\lambda_j^1 + \lambda_j^2 \geq \lambda_j^l, \quad j \in K \hspace{1cm} (P1e)$$

$$y_{l, A} \geq \sum_{j \in K} x_{i \rightarrow j, A}^l, \quad l \in \{1, 2\}; i \in N; A \subset N \hspace{1cm} (P1f)$$

$$\sum_{A \subset N} y_{l, A} = 1, \quad l \in \{1, 2\} \hspace{1cm} (P1g)$$

$$x_{i \rightarrow j, A}^l \geq 0, \quad j \in K; l \in \{1, 2\}; i \in N; A \subset N. \hspace{1cm} (P1h)$$
The objective (P1a) is the average packet delay of all queues over the entire network. (P1b) is the average packet delay of each UE group served by LTE as derived in (7). (P1c) is the average packet delay of each UE group served by LTE-U as derived in (8). (P1d) is the service rate of each divided UE group using the conservative model as given in (4). (P1e) is the total traffic constraint for each user type (6). (P1f) guarantees the consistency of the spectrum allocation as given in (2). (P1g) constrains the total bandwidth of each RAT to be one unit. (P1h) precludes negative bandwidth. The variables in P1 are $\mathbf{r}, \mathbf{x}, \mathbf{y}, \mathbf{\lambda}, \mathbf{t}$ which are the vector forms of $(r_{ij}^l)_{j \in \mathcal{K}, l \in \{1, 2\}}$, $(x_{i,A}^{l,j})_{j \in \mathcal{K}, j \in \mathcal{N}, A \subset \mathcal{N}, i \in \{1, 2\}}$, $(y_{l,A})_{A \subset \mathcal{N}, l \in \{1, 2\}}$, $(\lambda_{ji}^l)_{j \in \mathcal{K}, l \in \{1, 2\}}$, $(t_{ji}^l)_{j \in \mathcal{K}, l \in \{1, 2\}}$, respectively.

P1 is a bi-convex optimization problem because when variable $\mathbf{\lambda}$ (resp. $\mathbf{x}$) is fixed, all constraints are linear and the objective is a linear combination of convex functions in $\mathbf{x}$ (resp. $\mathbf{\lambda}$). Therefore, P1 can be solved by alternating optimization over $\mathbf{\lambda}$ and $\mathbf{x}$. Since the sequence of objective values obtained at each iteration is lower-bounded and non-increasing, the objective converges although it may not be a global minimum. As shown in Section VI, the alternating method achieves good performance.

**Theorem 1:** The (global) minimum average delay can be achieved by a sparse allocation, where the spectrum of each RAT is divided into at most $k$ segments. That is, there exists an optimal solution that satisfies

$$|\{A \mid y_{l,A} > 0, A \subset \mathcal{N}\}| \leq k, \quad \forall l \in \mathcal{M}$$

where $| \cdot |$ denotes the cardinality of a set. The same applies to the delay achieved by the alternating optimization method.

Theorem 1 is proved in Appendix B. The theorem guarantees that although the number of all possible patterns grows exponentially with the number of APs in the network, using a small number of patterns achieves the optimal performance.

**Theorem 2:** Assume the channel gains are jointly continuous random variables. The (global) minimum average delay can be achieved with at most $n - 1$ user groups served by multiple APs on each RAT. That is, there exists an optimal solution that satisfies

$$|\{j \mid x_{i_1,A_1}^{l,j} > 0, x_{i_2,A_2}^{l,j} > 0, \text{ for some } i_1, i_2 \in \mathcal{N} \text{ and } A_1, A_2 \subset \mathcal{N}\}| \leq n - 1, \quad \forall l \in \mathcal{M}.$$  

The same applies to the delay achieved by the alternating optimization method.
Theorem 2 is proved in Appendix C. The theorem guarantees that although we allow a UE group to be served by multiple APs, most UE groups will be associated with only one AP in the optimal solution.

IV. A Utilization-Dependent Allocation Scheme

In this section, we adopt the utilization-dependent service rate model (5). Unlike in the conservative scheme, we now let each AP adapt its transmission rates to the instantaneous set of active interfering APs.

In a stable interactive queueing system, each AP transmits (over each RAT) for a fraction of the time, referred to as the utilization. Denote the utilization of AP \( i \) over RAT \( l \) as \( \rho_i^l \in [0, 1] \). The analysis of interactive queueing system is difficult. As an approximation, we assume that different APs transmit independently over each RAT. The probability of an active interfering AP set \( I \) over RAT \( l \) is then:

\[
p_{l,I} = \left( \prod_{i \in I} \rho_i^l \right) \left( \prod_{i^{'}/\in I} (1 - \rho_i^l) \right).
\]

When UE group \( j \) is served, its instantaneous service rate is one of \( 2^n \) possible values depending on the set of active APs. Thus we use a certain “average” of \( 2^n \) independent M/M/1 queues to approximate the queueing behavior of UE group \( j \) with interactive queues. Such an approximation is reasonable under the premise that for any AP, the influence of other APs is adequately represented by their steady state probability distribution. Under each RAT, the average delay of UE group \( j \) is calculated as the expected delay over \( 2^n \) possible rates:

\[
t_1^j = \sum_{\mathcal{I} \subseteq \mathcal{N}} \frac{p_{1,I}}{(r_1^j, I - \lambda_1^j)^+}, \quad t_2^j = \sum_{\mathcal{I} \subseteq \mathcal{N}} \frac{p_{2,I}(2 + r_2^j, I \lambda_2^j \nu_j)}{2(r_2^j, I - \lambda_2^j)^+}.
\]

Similarly, the utilization of UE group \( j \) is calculated as its expected utilization over all possible sets of interfering APs:

\[
\sigma_{l}^j = \sum_{\mathcal{I} \subseteq \mathcal{N}} p_{l,I} \frac{\lambda_l^j}{r_{l,I}^j}, \quad l \in \{1, 2\}.
\]

Since AP \( i \) may serve multiple UE groups over each RAT, its utilization depends on the utilization of its associated UE groups. Therefore, the utilization of AP \( i \) over RAT \( l \) is approximated as
the average utilization over the spectrum used by AP $i$:

$$\rho_i^j = \frac{1}{\sum_{A:i \in A} y_{l,A}} \sum_{A:i \in A} \sum_{j \in K} \sigma_j^i x_{i,A}^j, \quad (14)$$

where $\sum_{A:i \in A} y_{l,A}$ is the total bandwidth used by AP $i$ over RAT $l$.

With the above approximation, the joint user association and utilization-dependent spectrum allocation problem is formulated as P2:

$$\begin{align*}
\text{minimize} & \quad U = \frac{1}{\sum_{j \in K} \lambda_j^1} \sum_{j \in K} (\lambda_j^1 t_1^j + \lambda_j^2 t_2^j) \quad (P2a) \\
\text{subject to} & \quad t_1^j = \sum_{\mathcal{I} \subset \mathcal{N}} \frac{p_{l,I}}{(r_{1,I}^j - \lambda_j^1)^+}, \quad j \in K \quad (P2b) \\
& \quad t_2^j = \sum_{\mathcal{I} \subset \mathcal{N}} \frac{p_{l,I}(2 + r_{1,I}^j \lambda_j^2 \nu)}{2(r_{2,I}^j \lambda_j^2)^+}, \quad j \in K \quad (P2c) \\
& \quad r_{1,I}^j = \sum_{A \in \mathcal{N}} \sum_{i \in \mathcal{K}} s_{i,A}^j x_{i,A}^j, \quad \mathcal{I} \subset \mathcal{N}; j \in K; l \in \{1, 2\} \quad (P2d) \\
& \quad \lambda_j^1 + \lambda_j^2 \geq \lambda_j^i, \quad j \in K \quad (P2e) \\
& \quad y_{l,A} \geq \sum_{j \in \mathcal{K}} x_{i,A}^j, \quad l \in \{1, 2\}; i \in \mathcal{N}; A \subset \mathcal{N} \quad (P2f) \\
& \quad \sum_{A \in \mathcal{N}} y_{l,A} = 1, \quad l \in \{1, 2\} \quad (P2g) \\
& \quad x_{i,A}^j \geq 0, \quad j \in K; l \in \{1, 2\}; i \in \mathcal{N}; A \subset \mathcal{N} \quad (P2h) \\
& \quad p_{l,I} = \left( \prod_{i \in \mathcal{I}} \rho_i^j \right) \left( \prod_{i \notin \mathcal{I}} (1 - \rho_i^j) \right), \quad l \in \{1, 2\}; \mathcal{I} \subset \mathcal{N} \quad (P2i) \\
& \quad \rho_i^j = \frac{1}{\sum_{A:i \in A} y_{l,A}} \sum_{A:i \in A} \sum_{j \in K} \sigma_j^i x_{i,A}^j, \quad l \in \{1, 2\}; i \in \mathcal{N} \quad (P2j) \\
& \quad \sigma_j^i \geq \sum_{I \subset \mathcal{N}} p_{l,I} \frac{\lambda_j^i}{r_{1,I}^j}, \quad l \in \{1, 2\}; j \in K. \quad (P2k)
\end{align*}$$

The objective (P2a) is also average packet delay of all queues of the entire network. (P2d) is the service rate of each divided UE group using utilization model as in (5). The variables in P2 are $r, x, y, \lambda, t, \sigma, \rho, p$ which are of the vector forms of $(r_{1,I}^j)_{\mathcal{I} \subset \mathcal{N}, j \in \mathcal{K}, l \in \{1, 2\}}$, $(x_{i,A}^j)_{j \in \mathcal{K}, i \in \mathcal{N}, A \subset \mathcal{N}, l \in \{1, 2\}}$, $(y_{l,A})_{A \subset \mathcal{N}, l \in \{1, 2\}}$, $(\lambda_j^i)_{j \in \mathcal{K}, l \in \{1, 2\}}$, $(t_l^j)_{j \in \mathcal{K}, l \in \{1, 2\}}$, $(\sigma_j^i)_{j \in \mathcal{K}, l \in \{1, 2\}}$, $(\rho_i^j)_{i \in \mathcal{N}, l \in \{1, 2\}}$, $(p_{l,I})_{\mathcal{I} \subset \mathcal{N}, l \in \{1, 2\}}$, respectively. P2 would be equivalent to P1 if we let $p_{l,N} = 1$. 

For every \( \mathcal{I} \neq \mathcal{N} \), and \( \sigma^j_l = 1 \) for every \( l \) and \( j \). Unlike P1, P2 is not bi-convex because new variables and nonlinear constraints are introduced. However, when fixing the variables \( \sigma, \rho \) and \( p \), it becomes bi-convex. Therefore, we divide P2 into two subproblems and solve them alternately to update all the variables.

Given \( \sigma, \rho \) and \( p \), we update \( r, x, y, \lambda, t \) by solving subproblem P3:

\[
\begin{align*}
\text{minimize} & \quad U = \frac{1}{\sum_{j \in K} I^j} \sum_{j \in K} (\lambda^j_1 t^j_1 + \lambda^j_2 t^j_2) \\
\text{subject to} & \quad \sum_{I \subseteq \mathcal{N}} p_{l,I} \lambda^j_l r^j_l \leq \sigma^j_l, \quad l \in \{1, 2\}; j \in \mathcal{K} \\
\end{align*}
\]

(P3b) constrains the utilization of UE groups. The structure of subproblem P3 is similar to that of P1. It can be solve by alternating optimization over \( \lambda \) and \( x \).

Given \( r, x, y, \lambda, t \), we update \( \sigma, \rho \) and \( p \). There are \( 2(2^n + n + k) \) variables and equations.

To solve them with low complexity, Algorithm 1 updates \( \sigma, \rho \) and \( p \) iteratively as in [15], [16]. Here \( m \) denotes the iteration. Convergence of the algorithm can be established similarly as in [15], [16]. In addition, the solution obtained by Algorithm 1 is feasible for (P2i)-(P2k).

**Algorithm 1 Update \( \sigma, \rho \) and \( p \)**

**Step 1. Initialization:** set initial utilization of UE groups \( \sigma(0) \);

**Step 2. Update:**

\[
\begin{align*}
\sigma^j_l(m) &= \sum_{\mathcal{I} \subseteq \mathcal{N}} p_{l,I}(m-1) \frac{\lambda^j_l}{r^j_l} , \quad l \in \{1, 2\}; j \in \mathcal{K} \\
\rho^j_i(m) &= \frac{1}{y_i} \sum_{\mathcal{A} \subseteq \mathcal{A}} \sum_{j \in \mathcal{K}} \sigma^j_l(m) x_{l,A} , \quad l \in \{1, 2\}; i \in \mathcal{N} \\
p_{l,I}(m) &= \left( \prod_{i \in \mathcal{I}} \rho^j_i(m) \right) \left( \prod_{i \notin \mathcal{I}} (1 - \rho^j_i(m)) \right) , \quad l \in \{1, 2\}; \mathcal{I} \subseteq \mathcal{N} \\
m &\leftarrow m + 1; \\
\end{align*}
\]

**Step 3.** If \( \| \sigma(m) - \sigma(m-1) \| \geq \epsilon \), where \( \epsilon \) is a small number, go back to Step 2; otherwise stop and obtain

\[
\sigma^j_l = \sigma^j_l(m), \quad \rho^j_i = \rho^j_i(m), \quad p_{l,I} = p_{l,I}(m), \quad l \in \{1, 2\}; j \in \mathcal{K}; i \in \mathcal{N}; \mathcal{I} \subseteq \mathcal{N}.
\]

After solving two subproblems alternately, P2 can be solved using Algorithm 2. Each step in Algorithm 2 is an easier problem. Convergence of Algorithm 2 can also be similarly established as in [15], [16]. Therefore, P2 can be solved iteratively with low complexity.
Algorithm 2 Iterative algorithm for solving P2

**Initialization:** $x = 0; x' = 1; \sigma^l_j = 1, \forall j \in \mathcal{K}, l \in \{1, 2\}; \rho^l_i = 1, \forall i \in \mathcal{N}, l \in \{1, 2\}; p_{l,x} = 1$

if $\mathcal{I} = \mathcal{N}, p_{l,x} = 0$ otherwise.

while $||x - x'|| > \epsilon$ do

1. $x' \leftarrow x;$

2. Update $r, x, y, \lambda, t$ by solving P3;

3. Update $\sigma, \rho, p$ by using Algorithm 1.

end while

V. Extension

So far, the packet lengths are assumed to be i.i.d with exponential distribution, which results in exponentially distributed service time for each packet. In this section, we show that the proposed schemes and algorithms are also applicable to general packet length distributions as long as the first and second moments if the service time can be written in the form:

$$E[X] = \frac{\beta}{r},$$

$$E[X^2] = \frac{\eta}{r^2},$$

where $r$ is the service rate and $\beta$ and $\eta$ are positive numbers. A wide class of distributions have such characteristics. For example, $\beta = 1$ and $\eta = 2$ correspond to exponential service time; $\beta = 1$ and $\eta = 1$ correspond to constant service time. In general, the queues in the system are not M/M/1 queues. In contrast to (7) and (8), the formula for calculating the average packet delay of each UE group becomes:

$$\hat{t}_1^j = \left(\frac{1}{2} \eta - \beta^2\right) \lambda_1^j + \beta r_1^j,$$

$$\hat{t}_2^j = \left(\frac{1}{2} \eta - \beta^2\right) \lambda_2^j + \beta r_2^j + \frac{1}{2} \nu \lambda_2^j (r_2^j)^2,$$

Then the objective function in P1 is rewritten as:

$$\hat{U} = \sum_{j \in \mathcal{K}} (\lambda_1^j \hat{t}_1^j + \lambda_2^j \hat{t}_2^j).$$

**Proposition 2:** The new objective function (17) is bi-convex in $\lambda$ and $r$.  

Proposition 2 is proved in Appendix D. Due to Proposition 2, the techniques proposed in Section III and Section IV still apply to the problem of user association and spectrum allocation with general packet length distribution.

VI. NUMERICAL RESULTS

Simulations are carried out using the network topology depicted in Fig. 2. The HetNet is deployed over a $100 \times 200$ m$^2$ area, divided into a hexagonal grid. Five APs, denoted by triangles, are uniformly randomly dropped at the vertices of the hexagons and 15 UE groups denoted by the squares are randomly dropped at the centers of the hexagons. The spectral efficiency is calculated using (3) assuming $\alpha_1 = 1$ and $\alpha_2 = 1/2$ with a 30 dB cap on the receive SINR.

| TABLE I |
| --- |
| **PARAMETER CONFIGURATIONS.** |
| Parameters | Value/Function |
| AP transmit power | 30 dBm |
| Total bandwidth on each RAT | 10 MHz |
| Average packet length | 0.5 Mbits |
| AP to UE pathloss | $140.7 + 36.7 \log_{10}(R)$ |
(SINR greater than 30 dB is regarded as 30 dB). Other parameters used in the simulation are given in Table I, and are compliant with the LTE standard [24]. The results for actual packet delay with the utilization service rate model are obtained through a packet-level simulator, which adapts the transmission time of each packet to the instantaneous active interfering APs.
A. The Conservative Allocation Scheme

The spectrum allocation and UE-AP association according to the solution of P1 under different traffic loads are shown in Figs. 3 and 4. Each smaller rectangle indicates the spectrum allocation at the corresponding UE group. The different colors represent different patterns. The amount of spectrum resources allocated to each user group under each pattern is denoted by the size of the corresponding bar inside the rectangle. The normalized average packet arrival rate relative to the total (sum) average arrival rate of each user group as well as the intensity of interference
nearby from the unlicensed band are shown above the rectangle. For example, for the UE group on the top left, 19/28 of its traffic load is assigned to be served using LTE while the remaining 9/28 is to be served using LTE-U. There are three interference levels, referred to as “High”, “Medium” and “Low”, according to the intensity of the interference nearby from the unlicensed band. For concreteness, we let the corresponding second moments of the vacation duration in (8) be $\nu_H = 1, \nu_M = 0.01, \nu_L = 0.0025$. When the second moment $\nu_L = 0.0025$, the average vacation time is similar to the average packet service time, which is about 0.05 seconds/packet.

Each line segment joining an AP and a user group describes their association. The color bars on the right of each figure shows the actual spectrum partition into different patterns. Theorem 1 and Theorem 2 can be verified by counting the number of pieces in the partition and the number of rectangles connecting multiple triangles, respectively. For instance, in Fig. 3, the licensed band is partitioned into 7 segments and the unlicensed band is partitioned into 6 segments, both numbers of partitions are smaller than the number of UE groups. Moreover, there is only one UE group served by multiple BTSs in the unlicensed band.

Fig. 3 shows the spectrum allocation patterns for the licensed band and unlicensed band in a high traffic scenario. The licensed band is partitioned into 7 segments and the unlicensed band is partitioned into 6 segments. As we can see, the spectrum resources allocated to each user group is roughly proportional to the corresponding traffic demand. Since the licensed band alone cannot support all the UE groups, some UE groups with low traffic demand are served by using the unlicensed band only and most licensed spectrum is allocated to the UE groups with high traffic. In addition, even though some UE groups have high traffic demand (e.g., the top right one), more unlicensed spectrum is allocated to them because they do not have much nearby interference from other RATs in the unlicensed band. In contrast, some UE groups with low traffic demand (e.g., the one in the middle with arrival rate 2 packets/second) may also be allocated licensed spectrum if the nearby interference from the unlicensed band is severe.

Fig. 4 shows the spectrum allocation patterns for the licensed and unlicensed bands in a medium/low traffic scenario. The licensed band is partitioned into 6 segments and the unlicensed band is partitioned into 4 segments. Since the licensed band alone is almost enough to support all the UE groups, most of them are served by the licensed band, and only a few of the UE groups are served by the unlicensed band because of the low traffic demand and lower interference generated within that band. In addition, most spectrum is allocated to the UE groups with high
traffic.

B. The Utilization-Dependent Allocation Scheme

Clearly, the delay obtained by using the conservative allocation is an upper bound on the actual delay obtained based on the utilization-dependent model. To validate the utilization-dependent model, a packet-level simulator is used to compare the theoretical delay with the actual delay. Fig. 5 compares theoretical and actual delays and service times using the utilization-dependent model. The service transmission time of a packet is part of the packet sojourn time excluding the queueing delay. The horizontal axis is the average traffic load of each UE group. All curves in Fig. 5 are based on the same spectrum allocation. The utilization approximation is quite accurate with light traffic. When traffic loads are heavy, the approximation becomes coarser due to more interactions among APs. In addition, compared with delay, the service time increases much more slowly with traffic load, indicating that the spectrum is efficiently allocated to mitigate interference among APs.

Fig. 5. Analytic and simulated delay for the utilization model.
Fig. 6 compares the optimization objective, average packet sojourn time, for the conservative and the utilization-dependent allocation schemes with orthogonal and full-reuse allocations. As the traffic increases to 40 and 50 packets/s, respectively, full-reuse and orthogonal allocations fail to support all the UE groups, suggesting these are the maximum throughputs under these two allocation schemes, respectively. Hence, the proposed conservative and utilization-dependent allocation schemes achieve a significantly larger throughput region than the other schemes. In addition, both conservative and the utilization-dependent allocation schemes yield significant gain especially in the high traffic scenario. The reason is that the proposed spectrum allocation schemes are traffic-aware, i.e., in the low traffic regime, it is better to reuse most of the spectrum, while in the high traffic regime, it is better to adaptively allocate the spectrum across APs. Furthermore, the proposed schemes exploit the particular characteristics of the RATs. That is, UE groups receiving less external interference from the unlicensed band are more likely to be allocated spectrum from that band. Furthermore, the actual delay based on the conservative allocation is also given, which is between the theoretical delay using conservative allocation and
Fig. 7. Spectrum allocation patterns for different loads.
the actual delay using the utilization-dependent allocation scheme. The utilization-dependent allocation scheme always outperforms the conservative allocation scheme as it accurately accounts for dynamic inter-cell interference. In low traffic, the allocation based on the utilization-dependent model reduces the average delay by about 40% compared to the conservative allocation, and this improvement becomes smaller as traffic loads increase.

Fig. 7 shows the optimal allocations for the licensed and unlicensed bands under different traffic loads using the conservative and utilization-dependent allocation schemes. Rectangles represent frequency bands and solid ones are used by the corresponding APs. A stack of solid rectangles then represents a reuse pattern. The length of each rectangle is proportional to the fraction of the whole spectrum occupied by the corresponding pattern. In both Fig. 7a and Fig. 7b, when traffic loads are low, the allocation scheme using the utilization-dependent model tends to be full reuse, while as traffic loads increase, it tends to orthogonalize the spectrum. In comparison with the conservative allocation, which remains orthogonal under different traffic loads, the utilization-dependent allocation uses the spectrum more efficiently since it approximates interaction among APs more accurately.

VII. Conclusion

We have studied spectrum allocation in downlink HetNets with multiple RATs over different bands using the average packet sojourn time as the performance metric. In addition to the licensed band, a queueing model with vacation has been proposed to model the additional delay of the unlicensed band. Two optimization based schemes have been proposed and shown to be highly effective. A sparse allocation (in terms of both the number of spectrum segments and user association) achieves the optimal network utility. Simulation results show that the proposed allocation schemes yield user associations and spectrum allocations, which utilize the spectrum of each RAT more efficiently compared with the benchmark allocation schemes, namely, orthogonal frequency reuse and full-frequency-reuse. Ongoing work is to extend the optimization framework to give a resource allocation solution with complexity that scales with the size of the network.
To analyze the queueing model with vacation and nonexhaustive service, we follow a similar technique as in [23]. Denote the residual service time at time $t$ by $R(t)$, and the waiting and service times for the $i$-th packet as $W^i$ and $X^i$, respectively. We have,

$$E[X] = \frac{1}{r}. \quad (17)$$

Assuming the queue is stable, by the Pollaczek-Khinchine formula,

$$E[W] = \frac{E[R]}{1 - \rho}, \quad (18)$$

and the average packet delay is given by:

$$E[T] = E[W] + E[X] = \frac{E[R]}{1 - \rho} + \frac{1}{r}, \quad (19)$$

where $\rho = \frac{\lambda}{r}$. Denote $M(t)$ as the number of packets served up until time $t$. To calculate $E[R]$, we can see from Fig. 8,

$$E[R] = \lim_{t \to \infty} \frac{1}{t} \int_0^t R(\tau) d\tau \quad (20)$$

$$= \lim_{t \to \infty} \frac{1}{t} \sum_{i=1}^{M(t)} \left( \frac{(X^i)^2}{2} + \frac{(V^i)^2}{2} \right) \quad (21)$$

$$= \lim_{t \to \infty} \frac{M(t)}{2t} \sum_{i=1}^{M(t)} \frac{(X^i)^2 + (V^i)^2}{M(t)} \quad (22)$$

$$= \frac{\lambda}{2} \left( E[X^2] + E[V^2] \right) \quad (23)$$

$$= \frac{\lambda}{r^2} + \frac{\lambda \nu}{2}. \quad (24)$$

The equalities hold in the almost sure sense. Substituting (24) into (19), we obtain (7).

To prove the convexity of (7), note that (7) can be written as the addition of two parts,

$$t = \frac{1}{(r - \lambda)} + \frac{\nu}{2} \frac{r \lambda}{(r - \lambda)} \quad (25)$$

Since both parts are convex in the pair $(r, \lambda)$ on $\mathbb{R}^2$, the sum is also convex in the pair $(r, \lambda)$ on $\mathbb{R}^2$. ■
Fig. 8. Residual service time $R(t)$.

APPENDIX B

PROOF OF THEOREM 1

We first reformulate $P_1$ by a change of variables. The proof then follows a similar geometric argument as in [12].

Consider a reformulation of $P_1$ by replacing constraints (P1c), (P1e), and (P1g) with the following three constraints, introducing a new collection of variables $(z^{i \to j}_{l,A})$:

$$r^j_l = \sum_{A \subseteq N} \left( \sum_{i \in N} s^{i \to j}_{l,A} z^{i \to j}_{l,A} \right) y_{l,A}, \quad j \in \mathcal{K}; \ l \in \{1, 2\} \tag{P5a}$$

$$\sum_{j \in \mathcal{K}} z^{i \to j}_{l,A} \leq 1, \quad l \in \{1, 2\}; \ i \in N; \ A \subseteq N \tag{P5b}$$

$$z^{i \to j}_{l,A} \cdot y_{l,A} \geq 0, \quad j \in \mathcal{K}; \ l \in \{1, 2\}; \ i \in N; \ A \subseteq N. \tag{P5c}$$

The new problem, referred to as $P_5$, is equivalent to $P_1$. This is because the feasible set for the rate tuple $r$ remains the same, which is easy to see by regarding $z^{i \to j}_{l,A}$ as the fraction of spectrum under pattern $A$ that AP $i$ allocates to UE group $j$ over RAT $l$. If one solves $P_5$, the actual spectrum allocations can be recovered as $x^{i \to j}_{l,A} = y_{l,A} z^{i \to j}_{l,A}$.

We show that if $r^*_{l}$ is an optimal rate tuple of $P_1$, then we can attain each sub-tuple $r_{l}^{*}$ over RAT $l$ with a $k$-sparse $y_{l}$. For each RAT $l$, let us begin with a generally nonsparse optimal solution $y_{l}$, whose support is $S_{l}$ ($y_{l,A} = 0$ if $A \notin S_{l}$). Let us freeze the optimal $z^{i \to j}_{l,A}$ variables and define a $k$-vector $q_{l,A}$ for every $A \in S_{l}$ with its elements determined by $q_{l,A}^{i} = \sum_{i \in N} s^{i \to j}_{l,A} z^{i \to j}_{l,A}$. According
to (P5a), a convex combination of the vectors \((q_{l,A})_{A \in S_l}\) with \((y_{l,A})_{A \in S_l}\) as coefficients form the optimal rate tuple \(r_l\) over RAT \(l\). By Carathéodory’s Theorem [25], \(r_l\) can be represented as a convex combination of at most \(k+1\) of those vectors. Moreover, \(r_l\) must be on the boundary, not in the interior of the convex hull of \((q_{l,A})\), because otherwise the rate tuple can be increased in all dimensions, contradicting the optimality assumption. Thus, for each RAT \(l\), we can identify a vector \(y^*_{l,A}\) whose support is a subset of \(S_l\) with \(k\) or fewer elements, such that the optimal \(r_l\) is a convex combination of \((q_{l,A})\) with \(y^*_{l,A}\) as coefficients.

**APPENDIX C**

**PROOF OF THEOREM 2**

The proof follows an analogous proof in [5]. The KKT conditions for P1 for nonzero elements of \(x\) and \(y\) are:

\[
\frac{\partial U}{\partial r_{l,j}} s_{i \rightarrow j}^{l_A} - \mu_{i,A}^l = 0, \quad (26a)
\]

\[
\sum_{i \in \mathcal{N}} \mu_{i,A}^l = \xi_l, \quad (26b)
\]

where \(\mu_{i,A}^l\) and \(\xi_l\) are the Lagrange multipliers for constraints (P1e) and (P1f), respectively. Assuming UE group \(j\) is served by two APs \(i_1\) and \(i_2\) over RAT \(l\), define \(s_{i \rightarrow j}^{l_A} = \sum_{A \subset \mathcal{N}} s_{i \rightarrow j}^{l_A}\) and \(\mu_{l,i} = \sum_{A \subset \mathcal{N}} \mu_{i,A}^l\). According to (26), we have

\[
s_{i_1 \rightarrow j}^{l_A} = \mu_{i_1}^l, \quad s_{i_2 \rightarrow j}^{l_A} = \mu_{i_2}^l. \quad (27)
\]

Following the argument in [5], a bipartite graph representation is used as in [26]. For each RAT, denote the UE groups served by multiple APs and the corresponding APs as nodes. An edge between a UE group and an AP represents an association. It remains to show that the graph contains no loop. Suppose there is a loop in the bipartite graph as shown in Fig. 9. Then there exists a sequence of nodes of users \(j_1, \ldots, j_p\) and a sequence of APs \(i_1, \ldots, i_p\), where user \(j_q\) is connected with AP \(i_q\) for \(q = 1, \ldots, p\), user \(j_{q+1}\) is connected with AP \(i_q\) for \(q = 1, \ldots, p-1\), and user 1 is connected with AP \(p\). The nodes are distinct otherwise we can find a smaller loop with this property. According to (27), the loop implies:

\[
\frac{s_{i_1 \rightarrow j_1}^{l_A}}{s_{i_2 \rightarrow j_2}^{l_A}} = \frac{\mu_{i_1}^l}{\mu_{i_2}^l}, \quad \frac{s_{i_2 \rightarrow j_2}^{l_A}}{s_{i_2 \rightarrow j_2}^{l_A}} = \frac{\mu_{i_2}^l}{\mu_{i_2}^l}, \quad \ldots, \quad \frac{s_{i_p \rightarrow j_p}^{l_A}}{s_{i_p \rightarrow j_p}^{l_A}} = \frac{\mu_{i_p}^l}{\mu_{i_p}^l} = 1.
\]
Since $s_{i \rightarrow j}^l$ is the sum of spectral efficiencies over all patterns over RAT $l$, it is a random variable based on random topology. Therefore, \( s_{i_1 \rightarrow j_1}^l \cdot s_{i_2 \rightarrow j_2}^l \cdot \ldots \cdot s_{i_p \rightarrow j_p}^l = 1 \) is a zero probability event, which shows that w.p.1 there is no loop in the graph. Since there are $n$ APs, the largest possible BGR without a loop has $n - 1$ user nodes, which proves Theorem 2.

**APPENDIX D**

**PROOF OF PROPOSITION 2**

Taking the second derivatives of $\hat{\lambda}_1^j \hat{t}_1^j$ with respect to $\lambda_1^j$ and $r_1^j$, we have:

$$\frac{\partial^2 \lambda_1^j \hat{t}_1^j}{\partial \lambda_1^j \partial r_1^j} = \frac{\eta r_1^j}{(r_1^j - \beta \lambda_1^j)^3}, \quad (28a)$$

$$\frac{\partial^2 \lambda_1^j \hat{t}_1^j}{\partial r_1^j \partial r_1^j} = \frac{\eta \lambda_1^j}{\beta} \left( \frac{1}{(r_1^j - \beta \lambda_1^j)^3} - \frac{1}{(r_1^j)^3} \right) + \frac{2 \beta \lambda_1^j}{(r_1^j)^3}. \quad (28b)$$

When $\frac{1}{\lambda_1^j} > \frac{2}{r_1^j}$, in other words, when the queue is stable, the derivatives are positive, which means $\lambda_1^j \hat{t}_1^j$ is bi-convex in $\lambda_1^j$ and $r_1^j$. In addition, since $\lambda_2^j \hat{t}_2^j$ is similar to $\lambda_1^j \hat{t}_1^j$ except for an additional term $\frac{\nu^j (\lambda_2^j)^2 r_2^j}{2(r_2^j - \beta \lambda_2^j)^4}$, which is also bi-convex in $\lambda_2^j$ and $r_2^j$, $\lambda_2^j \hat{t}_2^j$ is also bi-convex in $\lambda_2^j$ and $r_2^j$. Since $\hat{U}$ is a linear combination of bi-convex functions, we conclude that (17) is bi-convex in $\lambda$ and $r$.

**REFERENCES**

[1] Z. Zhou, D. Guo and M. L. Honig, “Allocation of licensed and unlicensed spectrum in heterogeneous networks”, in *Proc. IEEE Australian Communication Theory Workshop*, Melbourne, VIC, Australia, 2016.

[2] A. Khandekar, N. Bhushan, T. Ji, and V. Vanghi, “LTE-Advanced: Heterogeneous networks,” in *European Wireless Conference*, pp. 978-982, 2010.
[3] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, “What will 5G be?,” IEEE J. Sel. Areas Commun., vol. 32, no. 6, pp. 1065-1082, 2014.
[4] Q. Kuang and W. Utschick, “Energy management in heterogeneous networks with cell activation, user association and interference coordination,” IEEE Trans. Wireless Commun., early access 2016.
[5] Q. Kuang, W. Utschick, and A. Dotzler, “Optimal joint user association and resource allocation in heterogeneous networks via sparsity pursuit,” http://arxiv.org/abs/1408.5091, IEEE Trans. Sig. Proc. to appear. 2016.
[6] R. Etkin, A. Parekh and D. Tse, “Spectrum sharing for unlicensed bands,” in IEEE Journal on Selected Areas in Communications, vol. 25, no. 3, pp. 517-528, April 2007.
[7] J. Huang, V. G. Subramanian, R. A. Agrawal, and R. Berry, “Joint scheduling and resource allocation in uplink OFDM systems for broadband wireless access networks,” IEEE J. Sel. Areas Commun., vol. 27, pp. 226-234, Feb. 2009.
[8] D. Fooladivanda and C. Rosenberg, “Joint resource allocation and user association for heterogeneous wireless cellular networks,” IEEE Trans. Wireless Commun., vol. 12, no. 1, pp. 248-257, 2013.
[9] G. Lim, C. Xiong, L. J. Cimini Jr., and G. Y. Li, “Energy-efficient resource allocation for OFDMA-based multi-RAT networks,” IEEE Trans. Wireless Commun., vol. 13, no. 5, pp. 2696-2705, 2014.
[10] K. Shen and W. Yu, “Distributed pricing-based user association for downlink heterogeneous cellular networks,” IEEE J. Sel. Areas Commun., vol. 32, no. 6, pp. 1100-1113, 2014.
[11] S. Deb, P. Monogioudis, J. Miernik, and J. P. Seymour, “Algorithms for enhanced inter-cell interference coordination (eICIC) in LTE HetNets,” IEEE/ACM Trans. Netw., vol. 22, no. 1, pp. 137-150, 2014.
[12] B. Zhuang, D. Guo, and M. L. Honig, “Traffic-driven spectrum allocation in heterogeneous networks,” IEEE J. Sel. Areas Commun. Special Issue on Recent Advances in Heterogeneous Cellular Networks, vol. 33, pp. 2027-2038, 2015.
[13] B. Zhuang, D. Guo, and M. L. Honig, “Energy-efficient cell activation, user association, and spectrum allocation in heterogeneous networks,” IEEE J. Sel. Areas Commun. Special Issue on Energy-Efficient Techniques for 5G Wireless Communication Systems, early access 2016.
[14] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson and T. L. Marzetta, “Cell-free massive MIMO versus small cells,” http://arxiv.org/abs/1602.08232, 2016.
[15] F. Teng and D. Guo, “Resource management in 5G: A tale of two timescales,” in Proc. Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, 2015.
[16] F. Teng and D. Guo, “Dual-timescale spectrum management in wireless heterogeneous networks,” http://arxiv.org/abs/1604.02781, 2016.
[17] C. Godsil and G. Royle, Algebraic Graph Theory. New York: Springer-Verlag, 2001.
[18] N. Jindal and D. Breslin, “LTE and Wi-Fi in unlicensed spectrum: a coexistence study,” White paper, 2015.
[19] R. Ratasuk, M. A. Uusitalo, N. Mangalvedhe, A. Sorri, S. Iraji, C. Wijting, and A. Ghosh, “License-exempt LTE deployment in heterogeneous network,” in Proc. 9th International Symposium on Wireless Communication Systems, Paris, France, Aug. 2012.
[20] F. S. Chaves, E. P. L. Almeida, R. D. Vieira, A. M. Cavalcante, F. M. Abinader Jr., S. Choudhury, and K. Doppler, “LTE UL power control for the improvement of LTE/Wi-Fi coexistence.” IEEE Vehicular Technology Conference, pp. 1-6, 2013.
[21] E. P. L. Almeida, A. M. Cavalcante, R. C. D. Paiva, F. S. Chaves, F. M. Abinader Jr., R. D. Vieira, S. Choudhury, E. Tuomaala, and K. Doppler, “Enabling LTE/WiFi coexistence by LTE blank subframe allocation,” IEEE International Communications Conference, pp. 1-6, 2013.
[22] T. Nihtila, V. Tykhomyrov, O. Alanen, M. Uusitalo, A. Sorri, M. Moisio, S. Iraji, R. Ratasuk, and N. Mangalvedhe, “System
performance of LTE and IEEE 802.11 coexisting on a shared frequency band,” in Proc. IEEE Wireless Communications and Networking Conf., Shanghai, China, pp. 1056-1061, Apr. 2013.

[23] N. Tian and Z. G. Zhang, Vacation queueing models: theory and applications. Springer, 2006.

[24] 3GPP TR 36.814, “Further advancements for E-UTRA physical layer aspects (Release 9),” V9.0.0, March 2010.

[25] H. G. Eggleston, Convexity. No. 47, Cambridge University Press Archive, 1958.

[26] V. Gajic, J. Huang, and B. Rimoldi, “Competition of wireless providers for atomic users: Equilibrium and social optimality,” in Proc. Allerton Conf. Commun., Control, & Computing, pp. 1203-1210, Monticello, IL, USA, 2009.