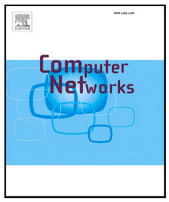
Computer Networks 189 (2021) 107893

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet



A reinforcement learning approach to queue-aware scheduling in full-duplex wireless networks

Hassan Fawaz [a,*], Melhem El Helou [a], Samer Lahoud [a], Kinda Khawam [b]

[a] Ecole Supérieure d'Ingénieurs de Beyrouth, Saint Joseph University of Beirut, Beirut, Lebanon
[b] Université Paris-Saclay, UVSQ, 78035, Versailles, France

ABSTRACT

Full-duplex communications promise to double the throughput of a wireless network, so long as the resulting interferences can be combated. Nonetheless, already dealing with the intricacy of determining base station-to-user radio channels, full duplex wireless networks need additional information on the channels in between all the user equipment. This information is necessary to determine user radio conditions, and thereafter efficiently allocate resources. A signaling overhead is likely to burden the user equipment, which already lack any methods to estimate and convey such user-to-user channel states. In this paper, we aim to circumvent the complexities and requirements of traditional scheduling techniques by introducing a reinforcement learning based scheduling algorithm for full-duplex wireless networks. This scheduling approach does not need to estimate user-to-user channels, and rather learns how to best allocate the network's radio resources. We show that our proposal can match scheduling with complete channel state information in terms of user equipment throughput, and that it performs well under multiple testing scheduling scenarios: increased user equipment numbers, randomized user equipment demand, and user equipment clustering among others.

1. Introduction

With an ever increasing global mobile data demand, already on the premises of 11 exabytes a month, and with septuple the traffic expected by the year 2022 [1], the declining efficiency of current half-duplex (HD) wireless networks is bound to pose a serious problem in bandwidth availability. HD wireless networks allocate a radio resource exclusively to one user equipment (UE) either for transmission or reception. They necessitate orthogonal time or radio channels for bidirectional transmissions. At most, only half of the bandwidth potential is being met. However, full-duplex (FD) communications, made possible due to the recent introduction of self-interference cancellation (SIC) technologies, are capable of exploiting the bandwidth in its entirety. In such networks, concurrent transmission and reception occurs on the same frequency band allowing, at least in theory, the duplication of current network capacity.

In our work, we consider an FD orthogonal frequency division multiple access (FD-OFDMA) network. This network exhibits an FD base station (BS) and HD UEs. This reduces interference problems, and keeps most of the complexities of implementing FD away from the terminals. The BS, being the FD device, transmits and receives simultaneously on the radio resources. The HD UEs form uplink–downlink pairs which share the same radio resources with one UE transmitting, and the other receiving.

This mode of operation incurs two paramount types of interferences. The first, self-interference, is the interference imposed by the transmitted signal from a full-duplex device, typically multiple times larger, on the received signal. This phenomenon degrades the performance of uplink UEs in the network. Second, with uplink and downlink UEs using the same radio resources, FD networks would also exhibit intra-cell co-channel interference. The signal from an uplink UE, transmitting with relatively high power, will interfere on the signal being received by its paired downlink UE, degrading thus its performance.

While self-interference can be battled with cancellation technologies [2], and the effects of intra-cell co-channel interference limited via scheduling techniques, there is still no evident manner in which FD networks can tackle the lack of complete channel state information (CSI). In order to properly schedule and distribute resources among pairs of uplink–downlink UEs, the network needs exact information on the channels between all the UEs, in addition to all BS-to-UE channels. In a single small cell network of only 5 uplink and 5 downlink UEs, the BS would have to be continuously updated with information on up to 35 radio channels. Ten of which are of type UE-to-BS and 25 of type UE-to-UE. A number that would dramatically increase in large cell scenarios. Current wireless networks standards do not count for UE-to-UE channels, nor do they include any protocols that permit estimating them. Our aim in this paper is to propose an algorithm capable of efficiently allocating the network resources without any knowledge of these inter-UE channels.

In FD-OFDMA wireless networks, the resources are allocated to pairs of uplink–dowlink UEs. In this paper, we propose a reinforcement learning based scheduling algorithm that aims to find the allocation

* Corresponding author.
E-mail address: hassan1fawaz@gmail.com (H. Fawaz).

https://doi.org/10.1016/j.comnet.2021.107893
Received 4 September 2020; Received in revised form 18 December 2020; Accepted 28 January 2021
Available online 3 February 2021

scheme which maximizes UE and network throughput. Our main goal is to avoid the added intricacies stemming from FD wireless communications, especially when it comes to signaling UE-to-UE channel states.

We prove that our proposed learning algorithm can match the efficiency of scheduling with complete CSI, and we assess its performance under different scheduling scenarios. We show the throughput gains of the algorithm in both full and dynamic buffer scenarios, and test its limits in the cases of low SIC, UE clustering and UE mobility. In what follows, we highlight our main contributions:

(a) We introduce a reinforcement learning based scheduling algorithm for FD-OFDMA wireless networks. This algorithm couples between UE pairs on the available network radio resources, and seeks to learn the allocation method which increases UE and network throughput. We explain the main challenges facing a machine learning scheduling algorithm, and highlight how we tackle them.
(b) We aim to compare the performance of our algorithm with the case where complete system information is available to the scheduler. To this end, we implement the FD Maximum Sum-Rate algorithm proposed in [3] on our system model. Since it is meant to work in a full-buffer scenario, we adapt it to our dynamic arrivals model. This algorithm is chosen because it has the same objective as our proposal, to maximize UE throughput. It is as such suitable as a reference to how we expect our algorithm to perform. Furthermore, we implement an HD scheduling algorithm with the same objective.
(c) We make our algorithm adaptable to the constantly varying UE radio conditions and buffer statuses, and show via simulations that it can closely emulate the performance of scheduling with complete CSI.
(d) Additionally, we test the validity of our algorithm against different scheduling challenges: Increased network UEs, randomized UE traffic, low SIC, and in the presence of UE clusters. We show that our algorithm manages to remain profitable regardless of the aforementioned factors.

The rest of this paper is structured as follows. Section 2 discusses the related works and our contributions. Section 3 has the system model: the radio model is detailed in Section 3.1, the traffic model in Section 3.2, and information on the channel states is presented in Section 3.3. We discuss reinforcement learning in Section 4, and we present our proposal in Section 4.1. The main challenges facing our proposed reinforcement learning algorithm are discussed in Section 4.2. Different simulation scenarios and results are presented in Section 5, wherein we discuss the performance of our algorithm under different scheduling conditions. This paper is concluded by Section 6.

2. Related works

Heretofore considered impossible due to interference problems [4], the rather recent introduction of self-interference cancellation techniques has spurred research into FD wireless networks, producing thus, an ample and rigorous state-of-the-art. The latter can be divided into two main categories. The first, a product of early stage research into FD networks, is concerned mainly with validating FD technologies and predicting their efficiency. The second, building on the thereafter verified gains of FD communications, is centered on proposing scheduling and power allocation algorithms for FD wireless networks.

The works in [5–8] revolve around assessing the possible gains of FD wireless networks. Their authors study the limitations and obstacles of implementing FD wireless communications. In one of the earliest works on in-band FD for wireless networks, the authors in [5] surveyed a range of SIC techniques and touched on the main challenges facing FD wireless networks. Aiming towards FD inclusion in upcoming 5G protocols, the authors in [6] proposed an FD module with which they simulated two types of FD networks: one when only the BS is FD capable, and the other when both the UEs and the BS are FD capable. Consequently, they assert the gains achievable from FD communications. In [7], different scenarios and implementations of possible FD wireless networks are discussed. Mainly, four representative application scenarios are presented: FD-MIMO networks, FD cooperative networks, FD-OFDMA cellular networks, and FD heterogeneous networks. Again, the authors use resource management problems for the purpose of validating wireless FD communications. With a more practical approach, the authors in [8] introduce a realistic model of a compact FD receiver. With this model at hand, the authors demonstrate via numerical evaluations the capacity gains of FD wireless networks, and bring insights onto the impact of SIC on the performance of these networks. Finally, as self-interference remains to be the main threat against the success of FD wireless communications, researchers continue to study the efficiency and effectiveness of SIC technologies. Authors of more recent works in [9–12] keep a close eye on the ongoing progress of interference cancellation techniques and their immediate relation to the advancement of FD technologies. These articles, and several more in the state-of-the-art, prove that FD wireless communications are profitable so long as the resulting interferences can be contained.

With FD wireless communications being well motivated, contributors to the FD wireless networks state-of-the-art steered research toward devising scheduling and power allocation algorithms. The authors in [13] design an optimal problem for joint power and scheduling in what they describe as a multi-carrier non-orthogonal multiple access system. They then propose a heuristic solution to avoid the complexity of their initial proposition. Similarly, but for FD-OFDMA systems, the authors in [14–22] put forward power allocation and scheduling schemes. They propose optimization problems with greedy objectives focused on sum-rate maximization. The joint task of scheduling and power allocation belongs to the category of mixed integer non-linear programming with exponential complexity and computational intractability. As such, the authors work on heuristic solutions which can produce near optimal performances, albeit bearing less complexity. Building upon our colleagues' work, we presented greedy scheduling algorithms for FD networks in [23] and [24]. We also proposed a more fairness oriented FD scheduler coupled with a power allocation scheme in [25], and additionally studied the effect of imperfect CSI on the performance of FD wireless networks in [26].

Multiple articles in the state-of-the-art have previously addressed utilizing machine learning to tackle intricate scheduling tasks. The authors in [27] propose a learning based approach to address multiple cellular network challenges such as limited data availability and convoluted sample data. The papers in [28] and [29] propose using deep learning to schedule resources in half-duplex wireless networks and allocate power in full-duplex ones, respectively. The authors in [30] present a reinforcement learning algorithm for radio resource control in half-duplex 5G vehicular networks.

These and countless others applications of machine learning in wireless networks exist in the related works as detailed in [31]. Nonetheless, non of the state-of-the-art approaches tackle the task of allocating time–frequency resources in an FD network. As we highlighted in the introduction, and detail later on, such wireless networks are of particular nature and possess an articulate relationship between uplink and downlink transmissions. The latter are correlated and scheduling on the uplink and the downlink cannot be done independently as in typical HD networks, as we highlighted in our previous work [32]. To the extent of the our knowledge, this is the only work that proposes a machine learning approach to scheduling in FD wireless networks.

In this paper, we propose a reinforcement learning approach to scheduling in FD-OFDMA wireless networks. FD networks could generate significant profit in terms of UE performance when a cellular network has all the information, on all the UEs, and all the corresponding radio channels. The UEs which least interfere upon each other are coupled on the radio resources they best perform on. A

scheduling objective, whether greedy or fair, can thereafter be applied. In such a scenario, the upper bound of doubling the capacity could be reached. Nonetheless, this optimal scenario is not feasible. Following our previous works, as well as the state-of-the-art, the intricacy of determining all the radio channels in an FD network could be daunting. Every UE would need information on the channel in between itself and all other UEs in proximity. Even if the scheduler was tasked with only determining a UE's strongest interferers, the UE in question would still have to regularly update the BS on multiple radio channels. This will inflict a signaling burden on the UE. The terminal, which we aim to alleviate from FD problems, would be overwhelmed with additional processing tasks. With different FD implementations, this problem only gets exponentially more difficult to tackle. For example, if the UEs are supposed to be FD as well, the number of interferers increases, and with it the processing and signaling needed to convey all the channel states to the scheduler. A different approach that does not assume perfect network state knowledge, could prove necessary for real life implementation of FD wireless networks.

Our approach to resource scheduling in FD-OFDMA wireless networks is centered on a machine learning algorithm. Specifically, an online reinforcement learning technique is used to associate UE pairs with the radio resources on which they can transmit/receive the most. Our proposition relies only on the feedback information currently provided by HD wireless networks: the channel quality indicator. Based on this information, the scheduler knows which coding and modulation schemes to use for downlink UEs. After the first resource allocation round, the scheduler uses the number of bits, transmitted or received, by a UE as a metric on the success of its scheduling decision. As time progresses, and with a reward system based on the number of bits transmitted and received, the algorithm continuously learns the allocation scheme which maximizes UE throughput.

Our reinforcement algorithm tackles the problem of scheduling in the presence of a non-full buffer traffic model. Non-full buffer traffic, like streaming and video, would make up to 78% of the global mobile traffic by the year 2021 [1]. This highlights the importance of studying how non-full buffer traffic affects scheduling in FD networks. Full-buffer models were used in the vast majority of state-of-the-art [13–20]. Owing mainly to their simplicity, the optimistic nature of full buffer models makes them attractive. Assuming that each UE has an infinite stream of bits in its buffer allows scheduling algorithms to produce expected results without accounting for real life wireless system aspects. For instance, the effect of multi-user diversity is exploited with full buffers. In addition, with all the UEs constantly requesting to transmit, a scheduling model cannot account for cases where interferences change because of UEs emptying their queues. This sets full buffer traffic models apart from reality, sometimes deceivingly anticipating positive results that might not exist in a real network. In the case of implementing a machine learning approach, as we detail in Section 4.2, accounting for non-full buffer traffic further complicates the scheduling task.

3. System model

3.1. Radio model

We consider a single-cell FD-OFDMA wireless network. This network exhibits a full-duplex BS and half-duplex UEs. The UEs are virtually divided into two sets: an uplink UE set, denoted by $\mathcal{U}$ and a downlink UE set, denoted by $\mathcal{D}$. The scheduler will pair between uplink and downlink UEs on any radio resource k of the set $\mathcal{K}$. This system is illustrated in Fig. 1. In our work, we assume that the physical layer is operated using an OFDMA structure. The radio resources are divided into time–frequency resource blocks. In the time domain, a resource block contains an integer number of OFDM symbols. In the frequency domain, a resource block contains adjacent narrow-band subcarriers and experiences flat fading. Scheduling decisions for downlink and uplink transmissions are made in every transmission time interval (TTI) t. At the beginning of each TTI, K resource blocks are to be allocated. The TTI duration is chosen to be smaller than the channel coherence time. With these assumptions, UE radio conditions will vary from one resource block to another, but remain constant over a TTI. The modulation and coding scheme (MCS), that can be assigned to a UE on a resource block, depends on its radio conditions. For performance evaluation, we consider in what follows LTE-like specifications, with a resource block being composed of 12 subcarriers and 7 OFDM symbols [33].

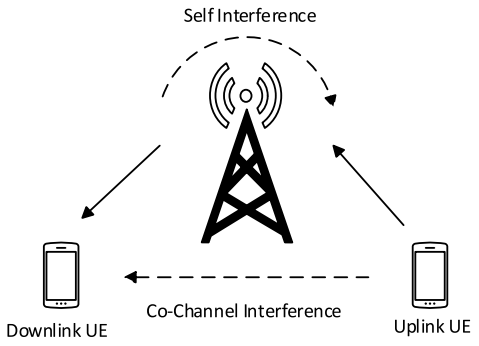


Fig. 1. Network model and interferences.

An adapted formula is used to calculate the SINR that takes into consideration the co-channel interference between a UE pair, and the self-interference cancellation performed by the BS. Let $P^u_{i,k}$ and $P^d_{j,k}$ denote the transmit power of the ith uplink user, and the transmit power of the BS serving downlink user j, respectively on the kth resource block. On resource block k, $h^u_{i,k}$ is the channel gain from the ith uplink user to the BS, and $h^d_{j,k}$ is the channel gain from the BS to the jth downlink user. Furthermore, $h_{ji,k}$ denotes the channel gain between the ith uplink user, and jth downlink user, on the kth resource block. Thus, $P^u_{i,k}|h_{ji,k}|^2$ is the co-channel interference on downlink UE j caused by uplink UE i, using the same resource block k. The self-interference cancellation level at the BS is denoted SIC. In particular, $\frac{P^d_{j,k}}{SIC}$ represents the residual self-interference power at the BS, on the kth resource block. Finally, $N_{0,k}$ and $N_{j,k}$ denote the noise powers at the BS and at the jth downlink user, respectively on the kth resource block. Eqs. (1) and (2) have the formulas for SINR calculation for uplink and downlink UEs respectively. For an uplink UE,

$$S^u_j(i,k) = \frac{P^u_{i,k}|h^u_{i,k}|^2}{N_{0,k} + \frac{P^d_{j,k}}{SIC}}, \quad i \in \mathcal{U}, \quad j \in \mathcal{D}. \tag{1}$$

For a downlink UE,

$$S^d_i(j,k) = \frac{P^d_{j,k}|h^d_{j,k}|^2}{N_{j,k} + P^u_{i,k}|h_{ji,k}|^2}, \quad i \in \mathcal{U}, \quad j \in \mathcal{D}, \tag{2}$$

where $S^u_j(i,k)$ is the SINR of uplink UE i on resource block k, while using the same resources as UE j. Similarly, $S^d_i(j,k)$ is the SINR of downlink UE j on resource block k, while being paired with UE i.

3.2. Traffic model

Our scheduling is queue-aware (Fig. 2). Each UE has a predefined throughput demand which determines the rate at which the UE will transmit or receive. A downlink UE has a queue at the BS, denoted Q^d_j, that it wants to receive. An uplink UE has a queue of bits it wants to transmit to the BS, denoted Q^u_i. UE queues are updated each TTI. They are filled according to a Poisson process with an arrival rate λ equal to the throughput demand. It is one of the most widely used and oldest traffic models [34]. Once the scheduling is done for a certain TTI,

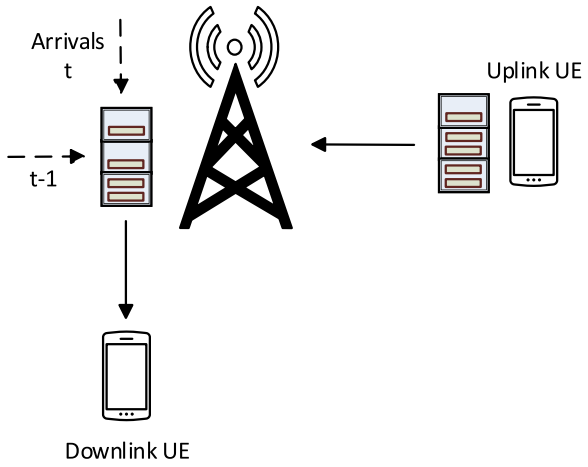


Fig. 2. Traffic model and UE queues.

the number of bits each UE can transmit or receive is calculated, and the UE queues are deducted accordingly. The traffic is packeted into small units known as transport blocks. Based on the MCS used and the number of resource blocks allocated for a UE, its transport block size is determined for the TTI. Any bits remaining in a UE queue at the end of a TTI are carried on to the next one. Our arrival model is configurable and different arrival processes can be implemented to account for different types of packets. Our main concern in this paper is that the arrivals are dynamic and that the traffic is non-full buffer, emulating thus real life traffic scenarios. In Algorithm 1, we illustrate how the UE queues are updated after a resource block is allocated. Q_x^u and Q_x^d represent the queue of a UE x on the uplink or downlink, respectively. T_{xjk}^u is the number of bits transmitted by a UE x on the uplink, and T_{ixk}^d is the number of bits received by a UE x on the downlink. A UE that has emptied its queue is removed from its corresponding set.

Algorithm 1 Queue Update Function

```
Update (x)
if x ∈ U
    Q_x^u ← Q_x^u − T_xjk^u
    if Q_x^u == 0
        U ← U − {x}
    end if
end if
if x ∈ D
    Q_x^d ← Q_x^d − T_ixk^d
    if Q_x^d == 0
        D ← D − {x}
    end if
end if
```

3.3. Channel state information

The state of a wireless channel is determined by the combined effect of several factors, the most pertinent being the path loss, the shadowing, and the fast fading. Knowledge of the channel on a wireless link permits adapting the transmission to the communication channel. This is essential for achieving reliable communications, and for making efficient resource allocation decisions.

Legacy HD networks would rely on feedback from the UEs to determine the current channel state [35]. These networks are concerned mainly with the channel in between the BS and the UEs, and different techniques are used to determine how often, and on which resource blocks, would this feedback information be required. The more periodic the feedback, the more accurate the channel estimation is.

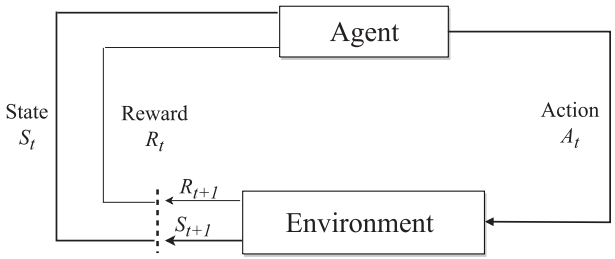


Fig. 3. Reinforcement learning model.

Full duplex communications add to the complexity of determining the CSI. In FD systems, additional information on the channel between the UEs of a pair is required. Not only do current wireless systems not count for such information, there is also no implemented method for which a UE can estimate the state of such UE-UE channels. Additionally, it is perceivable that continuously updating such information by the UEs would cause excessive overhead that they cannot handle.

We statistically model the inter-UE channel as follows:

$$h_{ji,k} = G_t G_r L_p A_s A_f \tag{3}$$

G_t and G_r are the antenna gains at the transmitter and the receiver, respectively. L_p represents the path loss between the two UEs, or equivalently the mean attenuation the signal undergoes on this channel. A_s and A_f are two random variables that respectively represent the shadowing effect, and the fast fading effect.

In our previous work in [26], we detailed the intricacies of scheduling without complete CSI knowledge, and the major losses that it would incur on FD gains. With every UE needing complete information on the channel in between itself and all other UEs, the number of inter-UE channels that need be estimated would grow by the order $\mathcal{O}(n^2)$, where n is the number of UEs in the network. Whilst research into reducing signaling overhead in current HD wireless networks is still ongoing, the estimation of n additional channels per UE, will drastically increase the signaling overhead. The BS might be able to cope with this added load, unlike the UEs which have limited processing capacity and battery life.

The most adequate solution to dealing with such massive expected overhead is to schedule the resources without this information at all. As such, we propose a reinforcement learning based scheduling algorithm for resource allocation in FD-OFDMA wireless networks. In our work, no information on any inter-UE channel is required prior to scheduling.

4. The reinforcement learning problem

In this subsection, we briefly explain the general reinforcement learning problem. Reinforcement learning is the idea of learning from interaction to achieve a goal [36]. The learner *i.e*, the decision maker in such a problem, is known as the agent. Everything else interacting with this agent is known as the environment. The environment and the agent interact at a sequence of discrete time steps, $t = 0, 1, 2, 3, \ldots$. At a moment in time t, the environment is in a state S_t. The agent takes an action A_t from the set of actions available in the current state $\mathcal{A}(S_t)$. As a consequence of the selected action, the agent will receive a reward $\mathcal{R}_{t+1}$, and subsequently, it will find itself in a new state S_{t+1}. This agent–environment interaction model is shown in Fig. 3.

Furthermore, the agent, in a state s, selects an action a with a probability p. This mapping is called the agent's policy, and is denoted π_t. $\pi_t(a|s)$ is thus the probability that $\mathcal{A}_t = a$ if $S_t = s$. As time progresses, a reinforcement learning algorithm should change its policy following the experience it has gained. The agent's goal when implementing new policies is to maximize the received rewards.

Reinforcement learning casts a wide net. Its framework is flexible and can be applied to different problems and via several ways. Relying on a dynamic iterative learning and decision-making process, in what follows, we use reinforcement learning to infer scheduling decisions in FD wireless networks in the absence of complete CSI. Our proposal can be likened to a multi-armed bandit problem [37]. We have a single state RL problem with the focus only being on deciding what action to take next (resource allocation).

4.1. Reinforcement learning scheduling algorithm

Let $p_{ijk}(t)$ be the probability that uplink UE i gets paired with downlink UE j on resource block k, during TTI t. The sum of all pairing probabilities is equal to 1, for each resource block k *i.e*, $\sum_{i\in\mathcal{U}}\sum_{j\in D} p_{ijk}(t) = 1\ \forall\ k \in \mathcal{K},\ \forall\ t$. Let k_{ij} be a binary value that is equal to one if pair (ij) is allocated resource block k, and 0 otherwise. As UE radio conditions vary from one resource block to another, learning scheduling decisions are done separately on each resource block. At $t = 0$, and at the beginning of the scheduling process, all possible UE pairings have equal probabilities of getting any resource. Among the different existing time-difference reinforcement learning algorithms, we choose to update the probabilities, after any resource block is allocated, as follows:

$$p_{ijk}(t+1) = \begin{cases} p_{ijk}(t) + \beta\mathcal{R}(1 - p_{ijk}(t)), & \text{if} \quad k_{ij} = 1 \\ p_{ijk}(t) - \beta\mathcal{R}p_{ijk}(t), & \text{otherwise,} \end{cases} \tag{4}$$

where β is the learning rate, chosen between 0 and 1, and $\mathcal{R}$ is the reward. The reward is evaluated as the number of bits the UE pair has transmitted/received (T_{ij}) on the allocated resource block, divided by the maximum number of bits (T_{max}) the UE pair could ideally send/receive over the RB *i.e*, if it was using the highest modulation order and coding rate. T_{max} is constant for all pairs. The reward is as such chosen in direct relation to a pair's radio conditions. The algorithm will reward UE pairs which best utilize the resource blocks, thus using the network bandwidth with utmost efficiency. Consequently, the more frequently a pair makes good use of the resource block, the more likely it is to get it within subsequent time slots. **P** is a 3-D matrix containing all the variables p_{ijk}.

The learning rate β controls both the speed with which the algorithm converges towards a largely preferred scheduling choice, as well as its efficiency. A high value of β would incur quicker decisions, but less efficient ones. Ideally, we want to chose the highest value of β that would always lead to a good scheduling decision with respect to maximizing UE throughput values.

In our problem, the agent is the scheduler at the BS. The environment is the UEs, as well as the resulting UE radio conditions after the pairing decisions. The reward is expressed in terms of bits transmitted by a UE pair on an allocated resource. Finally, the action is the process of selecting a UE pair to allocate resource blocks to.

4.2. Reinforcement learning scheduling challenges

Scheduling via a reinforcement learning algorithm poses several problems and challenges. In this section, we highlight these challenges and go over our approaches in tackling them.

4.2.1. Non-full buffer traffic

In our work, we focus on non-full buffer traffic scenarios. As such, any UEs which have emptied their queue within an allocation round, should be excluded from the scheduling process in the following one. UEs will be leaving and rejoining the network. Subsequently, all pairs which have an excluded UE should have their selection probability set to 0. Furthermore, when an excluded UE has new arrivals, it is re-factored into the scheduling task. With it are all the possible pairs within which this UE is contained. This raises another problem, what probabilities should these pairs take then?

We address this problem by introducing a temporary probability matrix **V** with pair-resource probability values v_{ijk}. At the beginning of each TTI, **V** will hold a copy of **P**. Resources are allocated within the current TTI according to the temporary matrix. After a resource block is allocated, **P** is updated following Eq. (4). When a UE empties its queue within a TTI, all pairs containing this UE have their probabilities in **V** set to 0. This insures that this UE will not get allocated anymore resources within the same TTI. The values of the remaining probabilities are normalized, *i.e*., each probability, for a certain resource block, is divided by the sum of the remaining probabilities. This keeps the sum of probabilities equal to one, unless of course all UE queues are empty. New arrivals are expected at the beginning of the new TTI (at least a limited number of bits). All UEs are now back in the scheduling process and the temporary matrix **V**, used for allocation, gets its values from the up-to-date **P**. The pseudo-code for the reinforcement learning algorithm is presented in Algorithm 1. Note that the total number of possible UE pairings $\mathcal{N}$ is defined as the number of uplink UEs multiplied by the number of downlink UEs.

Algorithm 2 RL Scheduling Algorithm

Requires: Set of states $\mathcal{S}$, actions $\mathcal{A}$, and rewards $\mathcal{R}$
Input: Learning rate $\beta \in [0,1]$
Initialize: $p_{ijk}(1) \leftarrow \frac{1}{\mathcal{N}},\ \forall\ (i,j,k) \in (\mathcal{U}\times D\times\mathcal{K})$
for TTI t=1,...,T
  $\mathbf{V} \leftarrow \mathbf{P}$
  for k=1,...,K
    Draw a pair according to the probabilities **V**
    Allocate k to the drawn UE pair $(i'j')$
    Compute $\mathcal{R} = T_{i'j'}/T_{max}$
    for $(i,j,k) \in (\mathcal{U}\times D\times\mathcal{K})$
      if $k_{ij} == 1$
        $p_{ijk}(t+1) \leftarrow p_{ijk}(t) + \beta\mathcal{R}(1 - p_{ijk}(t))$
      else
        $p_{ijk}(t+1) \leftarrow p_{ijk}(t) - \beta\mathcal{R}p_{ijk}(t)$
      end if
    end for
    if i' emptied its queue
      $v_{i'jk}(t) = 0,\ \forall\ (j,k) \in (D\times\mathcal{K})$
      Normalize **V**
    end if
    if j' emptied its queue
      $v_{ij'k}(t) = 0,\ \forall\ (i,k) \in (\mathcal{U}\times\mathcal{K})$
      Normalize **V**
    end if
  end for
end for

4.2.2. Exploration and exploitation

Our proposal makes it feasible to account for dynamic traffic. With UEs constantly leaving and joining back, the algorithm would not always select the same UE pair for any resource block. Every allocation the algorithm deems most suitable to maximize throughput values is only temporary, and bound to change once the pair(s) exits the allocation process or the radio conditions change. This makes our reinforcement learning algorithm similar to that of an ϵ-greedy one, where the algorithm will go into exploration with a probability ϵ [36]. However, in our case, the value of ϵ is determined by the demand of the UEs. For a low UE demand, ϵ is relatively high, and the algorithm could fall back into exploration several times within the same TTI. In the case of full buffer traffic, ϵ is equal to zero, and the algorithm would never go into exploration. Since we implement a non-full buffer traffic model, it is counter intuitive to manually assign a value for ϵ.

4.2.3. Online learning and dynamic radio conditions

As a result of shadowing and the time variant fast fading, any suitable pair selected by the learning algorithm to maximize UE and network throughput values, will not remain the best choice over the following TTIs. The radio conditions of each UE pair are bound to change from one TTI to another. In the case of non-full buffer traffic, this does not pose a major problem. After all, the algorithm is bound to regularly go into exploration mode. In all cases, our learning algorithm cannot be expected to find the relatively best allocation within one TTI. As such, it is correct to assume that the algorithm is learning the average radio conditions of the UE pairs across the TTIs, rather than the instantaneous ones. This is bound to be somewhat costly with respect to a greedy allocation method with full CSI.

The dynamics of the network imply that the UE pair that would maximize UE and network throughput on a certain resource block is constantly changing. We show, via our simulations, that our algorithm is capable of adapting to this change, as the allocation probabilities per resource block are updated each TTI.

4.2.4. Positive reinforcement

In our proposal, we always use a positive payoff. Unless a selected UE pair transmits zero bits, its probability of selection would always increase, no matter how slightly, for the next TTI. In the context of our simulation scenarios, we cannot determine if the number of bits a certain UE pair sent/received is good enough or not. A UE pair situated away from the BS might return a small reward, but it could still be among the best performing pairs in the current network. Using a negative payoff, *i.e.*, reducing the probability of selection for this pair, could in fact set the algorithm farther away from reaching its goal of maximizing UE and network throughput values.

4.2.5. Scheduling baseline

We seek to measure the efficiency and functionality of our proposal vs. a set of baseline scheduling techniques. In addition to the FD Max Sum-Rate algorithm in [3], we simulate other scheduling algorithms we have previously worked on. First, we use a random allocation scheme based on the popular round robin allocation technique. The algorithm, FD Round Robin, will aim to allocate resources in turn, and equally, among randomly generated pairs of UEs. No other factor is taken into consideration. Secondly, we simulate a more fairness oriented scheduling algorithm. It is made in reference to a baseline proportional fair algorithm. The main idea is to allocate resources to pairs of UEs based on their priorities. The priorities being a function of current and historic radio conditions. This algorithm can be seen in detail in our article in [26]. Finally, we note that the FD Max Sum-Rate algorithm was modified and adapted to our non-full buffer model. An optimal formulation of this algorithm can be seen in (5).

$$\text{Maximize} \quad \sum_{k\in\mathcal{K}}\sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{D}} z_{ijk}(R^u_j(i,k)+R^d_i(j,k)), \tag{5a}$$

$$\text{subject to} \quad \sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{D}} z_{ijk}\le 1,\ \forall k\in\mathcal{K}, \tag{5b}$$

$$\sum_{k\in\mathcal{K}}\sum_{j\in\mathcal{D}} z_{ijk}T^u_{ijk}\le D^u_i,\ \forall i\in\mathcal{U}, \tag{5c}$$

$$\sum_{k\in\mathcal{K}}\sum_{i\in\mathcal{U}} z_{ijk}T^d_{ijk}\le D^d_j,\ \forall j\in\mathcal{D}, \tag{5d}$$

$$z_{ijk}\in\{0,1\},\ \forall i\in\mathcal{U},\forall j\in\mathcal{D},\forall k\in\mathcal{K}. \tag{5e}$$

In the preceding problem z_{ijk} is a binary variable which is equal to one if uplink UE i is paired with downlink UEj on RB k, and zero otherwise. $R^u_j(i,k)$ is the rate of uplink UE i, paired with downlink UE j on RB k. Similarly, $R^d_i(j,k)$ is the rate of downlink UE j, paired with uplink UE i on RB k. T^u_{ijk} and T^d_{ijk} represent the number of bits UE i can transmit on RB k and the number of bits UE j can receive on RB k, respectively. D^u_i and D^d_j represent the demands of uplink UE i and downlink UE j, respectively. That is to say the number of bits in their queues.

Eq. (5a) is the objective of the optimization problem, to maximize the total sum-rate. Eq. (5b) indicates that an RB is allocated to at most one UE pair. The equations in (5c) and in (5d) are the buffer constraints. They verify that a UE is not allocated an RB it is not going to fully utilize.

Table 1
Simulation parameters.

Parameter	Value
Cell specifications	Single-Cell, 120, 500, 1000 m radius
Number of resource blocks	50
BS transmit power	24 dBm
Maximum UE transmit power	24 dBm
SIC value	10^{11} or 10^{9}
Number of UEs	5DL, 5UL or 10DL, 10UL
UE distribution	Uniform
Demand throughput	4 Mbps
Fast fading	Rayleigh. $\sigma = 1$
Shadowing	Normal law. $\mu = 0$ dB σ^2=10 dB
Path loss model	Extended Hata path loss model
TTI duration	1 ms

5. Simulation and results

We seek via our different simulation scenarios to address the validity and practicality of our machine learning scheduling proposal. First, and as the research into FD communications shifts from micro to macro cells, we assess the performance of our algorithm in a larger cell scenario. Second, we test the limits of our proposal, and show that with adequate parameters, it can match the performance of scheduling with complete CSI. Additionally, we test our algorithm under different circumstances: variable UE traffic, increased UE numbers, low self-interference cancellation values, and UE clustering among others.

The simulation parameters we used are presented in Table 1. The channel gain takes into account the path loss, the shadowing and the fast fading effects. The path loss is calculated using the extended Hata path loss model [38]. The shadowing is modeled by a log-normal random variable $A_s = 10^{(\frac{\xi}{10})}$, where ξ is a normal distributed random variable with zero mean and standard deviation equal to 10. The fast fading is modeled by an exponential random variable A_f with unit parameter. This model is used for urban zones, and it takes into account the effects of diffraction, reflection, and scattering caused by city structures. In our work, each simulation run serves a set of snapshots of the networks. Each snapshot has the UEs with different radio conditions on different resources.

In assessing the performance of our algorithm, we do not take into account the first few TTIs where the allocation process can be arbitrary. In Section 5.1 of the simulations the value of β is varied between 0.015 and 0.9 in order to study the significance of the learning rate. In the remainder of the simulations, the value of β is fixed at 0.015. This value of β guarantees the learning algorithm explores enough to find the pairs that maximize UE and network throughput every time.

5.1. Effect of the learning rate β

5.1.1. Case of small cell

We seek to study the effect of varying the learning rate on the performance of the algorithm. We consider a small cell of radius 120 m, the cell has 10 UEs: 5 uplink and 5 downlink. The throughput demand is 4 Mbps. The UE throughput values attained for β = 0.1, 0.3, 0.5, 0.7, and 0.9 are plotted in the cumulative distribution function (CDF) of Fig. 4.

For reference, a greedy FD Max Sum-Rate algorithm we enhanced is also plotted. This algorithm allocates resources to UE pairs that can

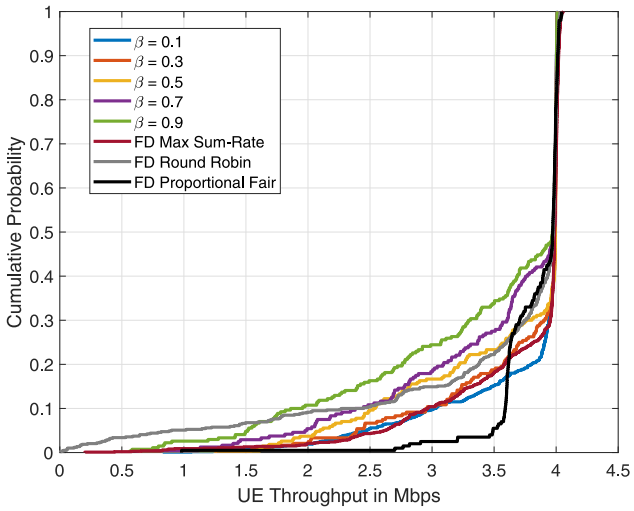


Fig. 4. Throughput as a function of the learning rate β, small cell.

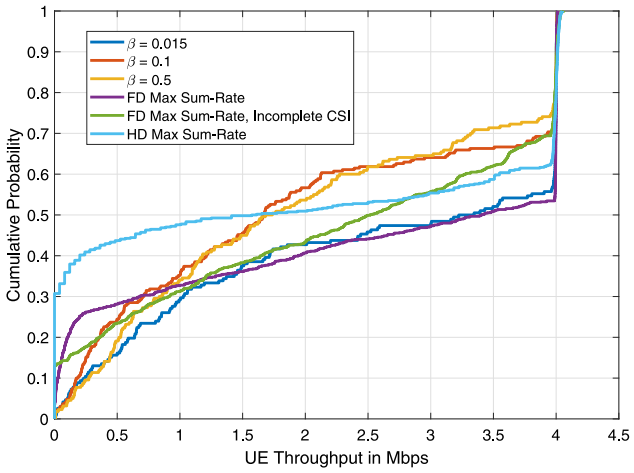


Fig. 5. UE throughput as a function of β. 500 m cell radius.

get the highest throughput values. This makes it an ideal reference to the performance we expect from an FD system which has complete channel knowledge. Additionally, we propose and plot a random resource allocation scheme in our FD Round Robin algorithm, and a fair allocation scheme in our FD Proportional Fair algorithm. For $\beta = 0.1$, around 70% of the learning algorithm UEs attained a throughput equal to their demand of 4 Mbps. This is almost identical to the FD Max Sum-Rate algorithm. As the value of β increases, the number of UEs attaining the nominal throughput value decreases. For $\beta = 0.9$, only about 50% of the simulated UEs attain a throughput equal to their demand. The resource allocation process becomes near random for this value of β, hence the similarity to the FD Round Robin algorithm. Moreover, the plot for FD Proportional Fair contrasts the difference in objectives with respect to our greedy algorithm. Only about 55% of the FD Proportional Fair algorithm UEs got a throughput equal to the demand of 4 Mbps, significantly lower than the 70% for our learning algorithm. However, for all except one of the Proportional fair UEs, the lowest recorded throughput value is 2.7 Mbps, significantly larger than 0.6 Mbps, the lowest recorded value for our learning algorithm (for $\beta = 0.1$). Our learning based algorithm is greedy and seeks to extract the utmost gain from the bandwidth, while FD Proportional Fair seeks to balance between bandwidth efficiency, and achieving fairness between the UEs.

Finally, we note that the lower the value of β, the more likely it is that the learning algorithm identifies the best pair to allocate each resource block to. Nonetheless, it would take longer for the algorithm to find this pair. That is to say that the higher the value of β is the quicker the algorithm can react to a change in the network, albeit at the cost of making more incorrect scheduling decisions, with respect to maximizing UE throughput values.

5.1.2. Case of large cell

Whilst FD communications are most suitable for small cells, the current state-of-the-art cancellation technologies allow mitigating self-interference by values upwards of 110 dB. This means that medium to large cell scenarios are pretty feasible. We repeat our simulation from the previous section, albeit with a cell radius of 500 m. This change in cell size, with the transmission powers being fixed, is bound to put more UEs in disadvantageous radio conditions. Cell edge UEs are more likely to have low SINR values. A bad scheduling decision is now more heavily punished. The results are shown in Fig. 5.

For $\beta = 0.1$, about 30% of the UEs attained a throughput equal to the demand. The FD Max Sum-Rate proposal attained a value close to 47%. Similar to before, the higher the value of β, the lower the performance of the algorithm. For $\beta = 0.5$, only about 25% of the UEs attained a throughput equal to the demand. Nonetheless, the gains with respect to HD wireless communications remain evident. HD Max Sum-Rate UEs has more UEs attaining a throughput equal to the demand, in comparison to our reinforcement learning proposal. Nonetheless, it also has about 30% of the UEs with zero throughput. Almost none of the service based learning UEs, regardless of the value of β, are denied throughput.

Additionally, we compare our machine learning solution to an FD Max Sum-Rate algorithm simulation done without any information on the UE-UE channels. (The method employed is illustrated in [26].) In such a case, almost 13% of the UEs were denied service, and 30% of the UEs attained a throughput equal to the demand. The performance of our algorithm for $\beta = 0.1$ thus barely outperforms scheduling with incomplete CSI. As such, we lower the value of the learning rate and simulate our learning algorithm for $\beta = 0.015$. In this case, our proposal can better match the performance of scheduling with complete CSI with about 44% of the UE attaining a throughput equal to the demand and less than 1% of them being denied throughput.

5.1.3. Selection of the value of the learning rate

In our aim to deduct the best value for β we simulate our algorithm for different values of the learning rate β and record how the algorithm would fair in terms of total network throughout with respect to scheduling with complete CSI, in the deterministic time slot. Fig. 6 has the results for $\beta = 0.015$, 0.1, and 0.5. In this simulation, and for the purpose of better distinguishing between the results, the radius of the cell is increased to 1 km. A wrong scheduling decision could now be more costly for the network.

For $\beta = 0.5$, the algorithm will reach an efficiency of about 75% in 200 TTIs. It no longer improves. For $\beta = 0.1$, the algorithm will take about 800 TTIs to reach an efficiency of 84% where it no longer improves on average. For $\beta = 0.015$, the algorithm is shown to be constantly improving. For this simulation scenario, it would reach upwards of 95% efficiency at around 4000 TTIs. A lower value of β would eventually lead to better efficiency, but at the cost of requiring more time to do so.

Since the value of $\beta = 0.1$ can barely outperform scheduling with incomplete CSI as illustrated in Fig. 5, a lower value of β is required for our macro-cell simulations. As such, for the remainder of the simulations, the value of β is set to 0.015.

5.2. Performance evaluation as a function of time

At the beginning of the simulation, the allocation process by our learning algorithm can be said to be arbitrary. Nonetheless, each TTI

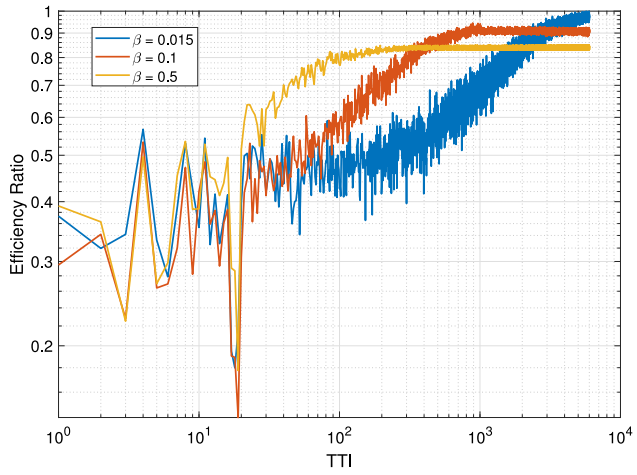


Fig. 6. Efficiency of the algorithm as a function of time and β.

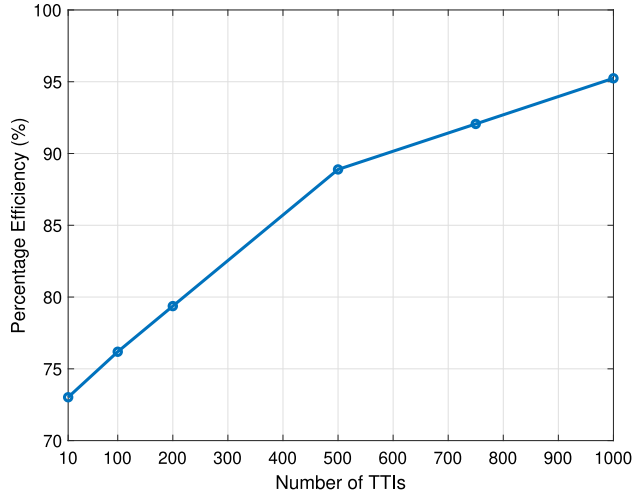


Fig. 7. Efficiency of the learning algorithm as a function of time.

the algorithm learns how to better allocate resources in a manner that maximizes UE and network throughput. For a 500 m cell radius, and a value of $\beta = 0.015$, we track the progress of our algorithm as a function of time.

We define the efficiency of the algorithm as the total network throughput attained by the learning algorithm divided by that attained by the FD Max Sum-Rate algorithm (with complete CSI). As explained before, we consider the latter to be a reference due to the similarity in objectives. Fig. 7 has a plot with the results. 10 TTIs, equivalently 10 ms, are enough for the algorithm to reach a 73% efficiency. At 1000 TTIs, or 1 s, the learning algorithm achieves about 95% of the throughput attained by FD Max Sum-Rate. From these results we can conclude that the learning algorithm can respond quickly to changes in the network, whether caused by dynamic radio conditions, UEs leaving or rejoining the network, or even UE mobility.

5.3. Effect of varying user characteristics

In this section, we vary different UE characteristics from randomizing traffic to clustering UEs and increasing UE numbers, and study the effects they have on UE performance. We show that regardless of the scenario at hand, our learning algorithm can mimic the performance of scheduling with complete CSI with high efficiency, and that it remains more profitable than scheduling without information on the inter-UE channels.

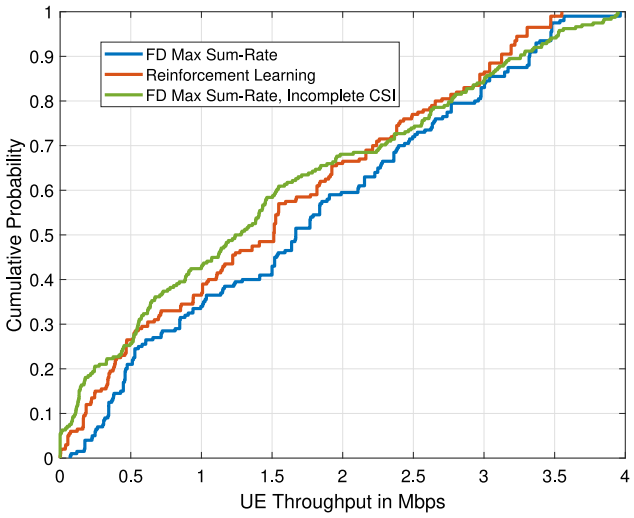


Fig. 8. Effect of randomized UE traffic.

5.3.1. Effect of randomized user demand

In this subsection, we aim to study the effect of different UE throughput demands on the performance of our reinforcement learning algorithm. To this end, we simulate the learning algorithm vs. FD Max Sum-Rate in a 500 m radius cell scenario with 10 UEs present. The throughput demand for each UE is set to random value uniformly chosen between 0 and 4 Mbps.

The performance of our reinforcement algorithm mimics that of greedy allocation with complete CSI, as illustrated in Fig. 8. Nonetheless, it is also evident that it lags in performance. The reinforcement learning algorithm would lose up to 9% in total network throughput in comparison. On the other hand, scheduling without inter-UE channel information costs 11% in network throughput efficiency and denies throughput to about 5% of the simulated UEs. This scenario is not very punishing to scheduling with incomplete CSI as many UEs have low throughput demands.

5.3.2. Performance assessment in the case of UE clustering

Additionally, we seek to study the effect of UE clustering on the performance of our algorithm. For a cell of 500 m radius, the UEs are all placed within 200 m distance from the BS. The SIC value is returned to the relatively good value of 10^{11}, and the remainder of the simulation parameters are left unchanged. Following the SINR calculation for downlink UEs in Eq. (2), the proximity of uplink and downlink UEs (as a result of UE clustering) degrades the radio conditions of downlink UEs. A wrong scheduling decision is bound to now have a more grievous effect on the performance in general, and on downlink UEs throughputs specifically.

Fig. 9 shows the CDF plots of the downlink UE throughput values for FD Max Sum-Rate scheduling with both complete and incomplete CSI, and that of our learning proposal as well. The complete CSI FD Max Sum-Rate scheduling algorithm edges out the learning algorithm in terms of UEs attaining a throughput equal to the demand (35% to 28%). The network would lose about 10% in total throughput. Nonetheless, this loss for the learning algorithm is mainly found in downlink UEs which on average deliver around 86% of the throughput attained by their FD Max Sum-Rate counterparts. Uplink UEs however almost match their counterparts with around 97% of the throughput. This form of UE clustering incurs a small performance penalty on our reinforcement learning scheduling proposal. Nonetheless, Max Sum-Rate scheduling with incomplete CSI incurs a higher loss. In this case, 18% of the simulated downlink UEs are denied any resources and the network loses 22% in terms of total throughput efficiency in comparison to scheduling with complete CSI.

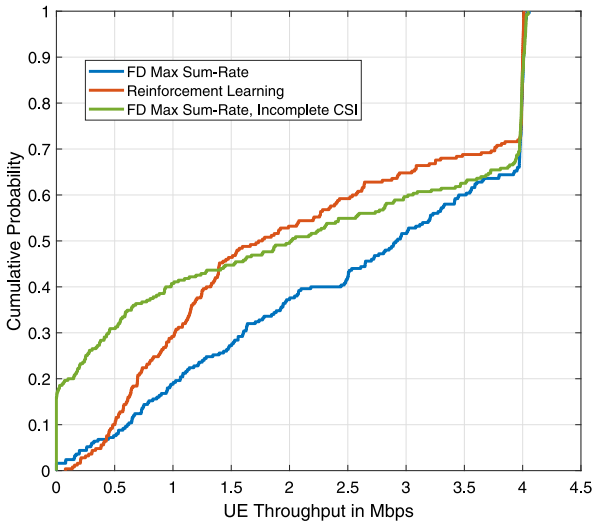


Fig. 9. Effect of clustering on downlink UE performance.

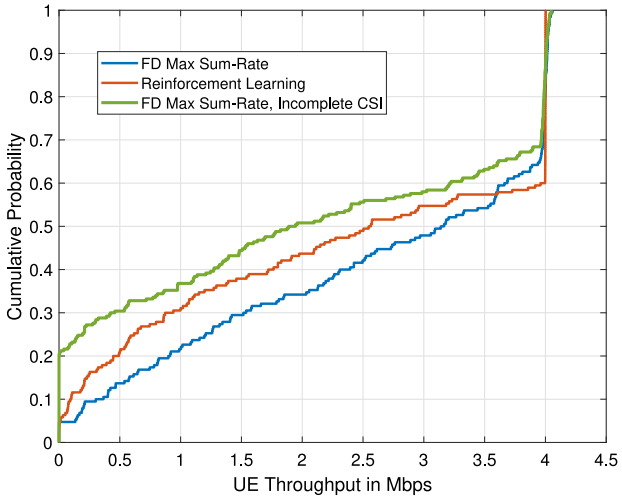


Fig. 10. Effect of UE mobility on performance.

5.3.3. *Effect of UE mobility on performance*

In this section, we study the effect of mobile UEs on the performance of our algorithm. We consider a random walk model [39] in determining the movement of the UEs. Each TTI, the UEs will move from a current location to a new one by choosing a speed and a direction randomly from the uniform intervals $[speedmin, speedmax]$ and $[0, 2\pi]$, respectively. The minimum and maximum speeds are chosen as the average velocity of a walking person (0.5 m/s) and the average velocity of a moving car (20 m/s), respectively. As the positions of the UEs change, their individual radio conditions will vary. This variation is related to their proximity to the BS, as well as to the resulting changes in the shadowing and the fast fading effects. We simulate our learning algorithm alongside FD Max Sum-Rate in both complete and incomplete CSI scenarios. The simulation is done with 10 UEs present in a 500 m radius cell. Fig. 10 has a CDF plot of the resulting UE throughputs.

Our algorithm shows more UEs attaining the throughput demand (40% to about 35%). Nonetheless, on average FD Max Sum-Rate UEs, scheduled with complete CSI, will get higher throughput values. It has a median UE throughput value equal to about 3.2 Mbps compared to a 2.6 Mbps median value for our learning algorithm. UE mobility will force erroneous decisions (with respect to maximizing UE throughput) and incite changes in performance for both algorithms. However, there is no noticeable added degradation in performance for our proposal in comparison to scheduling with complete CSI, and with respect to previous scheduling scenarios. Our learning algorithm can adapt over time to changes in UE radio conditions, limiting its losses to about 10% in total network throughput. In comparison, scheduling without information on the inter-UE channels causes higher losses in efficiency. Around 20% of the UEs in that case attain a throughput of 0 Mbps and the total efficiency of the algorithm drops to around 79%.

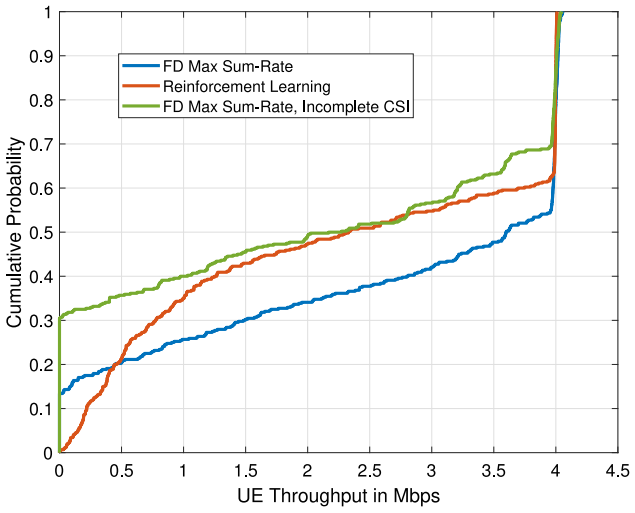


Fig. 11. UE throughput as a function of UE numbers.

5.3.4. *Effect of an increase in the number of UEs*

We seek to study the effect of increased UE numbers in the cell on the performance of our proposal. The number of UEs is increased to 20: 10 uplink and 10 downlink. The number of resource blocks is also doubled. Our aim is to study how the learning algorithm copes with increased scheduling options and not to increase the network load. The cell radius is 500 m and the SIC value remains at the relatively good value of 10^{11}. The throughput demand is 4 Mbps and the learning factor β is set to 0.015. Accordingly, there are 100 different possible pairing scenarios. We simulate the learning algorithm vs. the FD Max Sum-Rate proposal for both scenarios of complete and incomplete CSI.

Fig. 11 has the CDF plot of the corresponding UE throughputs. Around 45% of the FD Max Sum-Rate UEs, simulated with complete CSI, attained a throughput equal to the demand, compared to 38% for the learning algorithm. Around 14% of the former UEs were completely denied throughput compared to none in the case of the learning algorithm. Our proposal makes good enough scheduling decisions to mimic the FD Max Sum-Rate algorithm with complete CSI. The algorithm will lose a small part of the efficiency in terms of total network throughput. Arguably, this is a trade-off between efficiency and a more fair resource allocation. Furthermore, we simulate how the FD Max Sum-Rate algorithm would fair if the inter-UE channel information was not available. In such a case, almost 30% of the UEs are completely denied throughput. Additionally, every simulated UE attains a throughput value lower than that achieved by our learning proposal. In terms of total network throughput, our learning algorithm loses about 5 to 10% in efficiency when compared to FD Max Sum-Rate scheduling with complete CSI. Scheduling without complete CSI costs upwards of 22%.

5.3.5. *Scalability of the problem*

Furthermore, we look at how an increase in the number of UEs in the network affects the time needed for the algorithm to become efficient. Fig. 12 has a plot detailing the number of TTIs needed for our learning algorithm to reach 90% efficiency (with respect to FD Max Sum-Rate with complete CSI).

As the number of UEs in the network increases, more time is required to reach the 90% efficiency mark. At 20 UEs 3000 TTI are needed. At 60 UEs, about 8000 TTIs are needed. Nonetheless, with 1 TTI equaling 1 ms, the problem remains scalable even as the number of UEs in the network increases.

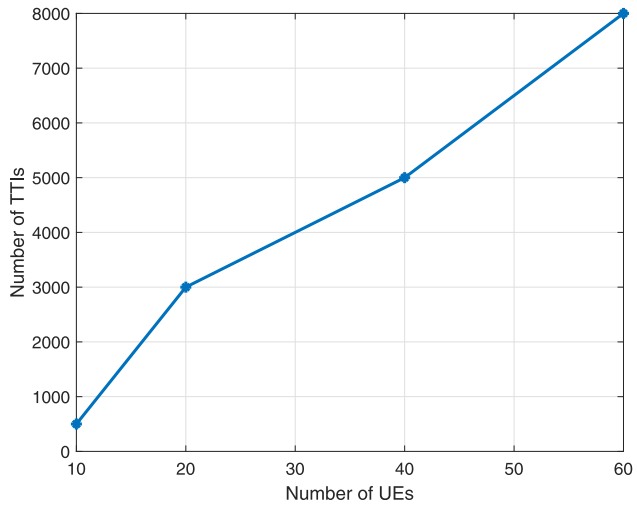


Fig. 12. Number of TTIs needed to attain 90% efficiency as a function of the number of UEs.

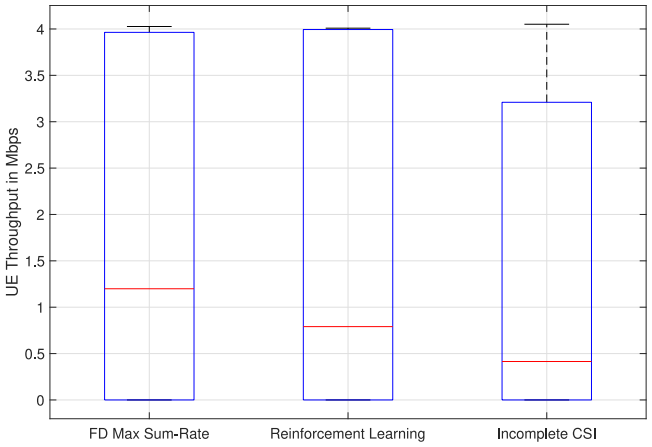


Fig. 13. Effect of low SIC on UE performance.

5.4. UE performance under low SIC

We lower the value of the SIC factor to 10^9. With 10 UEs present in a 500 m radius cell, the remainder of the simulation parameters remain unchanged as above. This is bound to negatively impact the performance of uplink UEs in the network as their SINR values degrade (see 5.3.1). Multiple uplink UEs, especially those on the cell borders, would now suffer from bad radio conditions. An incorrect scheduling decision, with respect to maximizing UE and network throughput, made by the learning algorithm would be more severely punished than before.

Fig. 13 has box plots [40] of the resulting UE throughput values for our reinforcement learning algorithm and FD Max Sum-Rate, both with and without complete CSI. Both our algorithm and Max Sum-Rate scheduling with complete CSI show again a similar distribution of UE throughputs with maximums equal to the demand, and minimums equal to zero. Nonetheless, the FD Max Sum-Rate algorithm with complete CSI has more UEs achieving a throughput equal to the demand, and fewer ones attaining zero throughput. The increased cost on scheduling errors incurred by lowering the SIC factor are visible on the network performance as a whole, where the learning algorithm would lose up to 19% in total network throughput with respect to scheduling with complete CSI. This is not a significant drop in performance. Nonetheless, uplink UEs are the most affected by this degradation. Downlink UEs scheduled by our learning algorithm attain throughput values, on average, equal to 97% of those achieved by their complete CSI FD Max Sum-Rate counterparts, but uplink UEs only manage around 60%. In comparison to scheduling with incomplete CSI, the median UE throughput value for our algorithm sits at about 0.75 Mbps. For the former it is about 0.4 Mbps. In addition, uplink UEs scheduled without information on the inter-UE channels manage only about 40% efficiency in comparison to scheduling with complete CSI.

Table 2
Efficiency with respect to scheduling with complete CSI.

	RL algorithm	Incomplete CSI
Randomized demand	91%	89%
Clustering	90%	78%
Mobility	90%	79%
Increased UE numbers	93%	78%
Low SIC	79%	68.5%

5.5. Comments on the results

In Table 2, we summarize the performance of our algorithm with respect to our simulations on scheduling with complete CSI and compare it to scheduling with incomplete CSI.

We highlight the different scenarios we considered to test the efficiency of our algorithm. They include randomized demand, UE clustering, UE Mobility, UE densification, as well as low self-interference cancellation capabilities. Our algorithm significantly outperforms scheduling with incomplete CSI in all cases. As the simulation environment gets more difficult (more devices, less resources, etc...), the better our proposal performs with respect to scheduling with incomplete CSI.

6. Conclusion

In this paper, we present a reinforcement learning based approach to scheduling in FD-OFDMA wireless networks. Our main objective is to avoid the added intricacies of scheduling in FD wireless networks. Specifically, we let go the unrealistic assumption of perfect CSI, as well as the regularly expected knowledge of all UE-to-UE channel states. Our algorithm is queue-aware and factors dynamic arrivals into account. We detail the main challenges facing a machine learning scheduling proposal, focusing on the effects of non-full buffer traffic and dynamic radio conditions on the performance of the algorithm. We test our proposal in multiple scheduling scenarios from randomized UE traffic, to UE clustering and in the presence of low SIC. While UE clustering degrades the performance of downlink UEs, and low SIC that of uplink UEs, we show that our learning proposal still performs well in terms of UE and network throughput. Furthermore, we show that in the case of mobile UEs, no added degradation in performance is incurred. We accordingly verify the validity of our proposal regardless of any obstacles facing the scheduling task. In our proposal, inter-cell interferences are neglected. We assumed that inter-cell interference coordination algorithms are in place. To fully assess the impact of the FD features and resulting interferences, we focused on a single-cell scenario. In future works, we will tackle multi-cell scenarios and the possible implementation of machine learning based scheduling with multi-cell coordination in a realistic setting.

CRediT authorship contribution statement

Hassan Fawaz: Conception and design of the study, The simulation of the proposed algorithms, Editing and formatting of the manuscript. **Melhem El Helou:** Conception and design of the study, The simulation of the proposed algorithms, Editing and formatting of the manuscript. **Samer Lahoud:** Conception and design of the study, The simulation of the proposed algorithms, Editing and formatting of the manuscript. **Kinda Khawam:** Conception and design of the study, The simulation of the proposed algorithms, Editing and formatting of the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors would like to acknowledge the National Council for Scientific Research of Lebanon (CNRS-L) and the Reasearch Council at Université Saint-Joseph de Beyrouth for granting a doctoral fellowship to Hassan Fawaz.

References

[1] C.V. Mobile, Cisco visual networking index: global mobile data traffic forecast update, 2017–2022 white paper, 2018.
[2] S. Hong, J. Brand, J.I. Choi, M. Jain, J. Mehlman, S. Katti, P. Levis, Applications of self-interference cancellation in 5g and beyond, IEEE Commun. Mag. 52 (2) (2014) 114–121.
[3] A.C. Cirik, K. Rikkinen, R. Wang, Y. Hua, Resource allocation in full-duplex ofdma systems with partial channel state information, in: International Conference on Signal and Information Processing (ChinaSIP), 2015 IEEE China Summit, IEEE, 2015, pp. 711–715.
[4] A. Goldsmith, Wireless communications, 2005.
[5] A. Sabharwal, P. Schniter, D. Guo, D.W. Bliss, S. Rangarajan, R. Wichman, In-band full-duplex wireless: Challenges and opportunities, IEEE J. Sel. Areas Commun. 32 (9) (2014) 1637–1652.
[6] M.G. Sarret, G. Berardinelli, N.H. Mahmood, P. Mogensen, Can full duplex boost throughput and delay of 5g ultra-dense small cell networks? in: 2016 IEEE 83rd Vehicular Technology Conference (VTC Spring), IEEE, 2016, pp. 1–5.
[7] L. Song, Y. Li, Z. Han, Resource allocation in full-duplex communications for future wireless networks, IEEE Wirel. Commun. 22 (4) (2015) 88–96.
[8] J. Marasevic, J. Zhou, H. Krishnaswamy, Y. Zhong, G. Zussman, Resource allocation and rate gains in practical full-duplex systems, IEEE/ACM Trans. Netw. 25 (1) (2017) 292–305.
[9] K.-C. Hsu, K.C.-J. Lin, H.-Y. Wei, Inter-client interference cancellation for full-duplex networks, in: IEEE INFOCOM 2017-IEEE Conference on Computer Communications, IEEE, 2017, pp. 1–9.
[10] T. Chen, M. Baraani Dastjerdi, J. Zhou, H. Krishnaswamy, G. Zussman, Wideband full-duplex wireless via frequency-domain equalization: Design and experimentation, in: The 25th Annual International Conference on Mobile Computing and Networking, 2019, pp. 1–16.
[11] N.V. Shende, Gürbüz, E. Erkip, Half-duplex or full-duplex communications: Degrees of freedom analysis under self-interference, IEEE Trans. Wireless Commun. 17 (2) (2018) 1081–1093.
[12] L. Irio, R. Oliveira, L. Oliveira, Characterization of the residual self-interference power in full-duplex wireless systems, in: 2018 IEEE International Symposium on Circuits and Systems (ISCAS), 2018, pp. 1–5.
[13] Y. Sun, D.W.K. Ng, Z. Ding, R. Schober, Optimal joint power and subcarrier allocation for full-duplex multicarrier non-orthogonal multiple access systems, IEEE Trans. Commun. 65 (3) (2017) 1077–1091.
[14] A.C. Cirik, K. Rikkinen, M. Latva-aho, Joint subcarrier and power allocation for sum-rate maximization in ofdma full-duplex systems, in: 2015 IEEE 81st Vehicular Technology Conference (VTC Spring), IEEE, 2015, pp. 1–5.
[15] J.M.B. da Silva, Y. Xu, G. Fodor, C. Fischione, Distributed spectral efficiency maximization in full-duplex cellular networks, in: 2016 IEEE International Conference on Communications Workshops (ICC), IEEE, 2016, pp. 80–86.
[16] S. Goyal, P. Liu, S.S. Panwar, R.A. DiFazio, R. Yang, E. Bala, Full duplex cellular systems: will doubling interference prevent doubling capacity? IEEE Commun. Mag. 53 (5) (2015) 121–127.
[17] C. Nam, C. Joo, S. Bahk, Joint subcarrier assignment and power allocation in full-duplex ofdma networks, IEEE Trans. Wireless Commun. 14 (6) (2015) 3108–3119.
[18] P. Tehrani, F. Lahouti, M. Zorzi, Resource allocation in ofdma networks with half-duplex and imperfect full-duplex users, in: 2016 IEEE International Conference on Communications (ICC), IEEE, 2016, pp. 1–6.
[19] B. Di, S. Bayat, L. Song, Y. Li, Z. Han, Joint user pairing, subchannel, and power allocation in full-duplex multi-user ofdma networks, IEEE Trans. Wireless Commun. 15 (12) (2016) 8260–8272.
[20] C. Nam, C. Joo, S. Bahk, Radio resource allocation with inter-node interference in full-duplex ofdma networks, in: 2015 IEEE International Conference on Communications (ICC), IEEE, 2015, pp. 3885–3890.
[21] W. Choi, H. Lim, A. Sabharwal, Power-controlled medium access control protocol for full-duplex wifi networks, IEEE Trans. Wireless Commun. 14 (7) (2015) 3601–3613.
[22] T. Chen, J. Diakonikolas, J. Ghaderi, G. Zussman, Hybrid scheduling in heterogeneous half-and full-duplex wireless networks, IEEE/ACM Trans. Netw. 28 (2) (2020) 764–777.
[23] H. Fawaz, S. Lahoud, M. El Helou, Queue-aware scheduling in full-duplex wireless networks, Wirel. Netw. (2020) 1–17.
[24] H. Fawaz, S. Lahoud, M.E. Helou, M. Ibrahim, Optimal max-sinr scheduling in full-duplex ofdma cellular networks with dynamic arrivals, in: 2018 Wireless Days (WD), 2018, pp. 196–201.
[25] H. Fawaz, S. Lahoud, M.E. Helou, J. Saad, Queue-aware priority based scheduling and power allocation in full-duplex ofdma cellular networks, in: 2018 25th International Conference on Telecommunications (ICT), 2018, pp. 15–20.
[26] H. Fawaz, S. Lahoud, M.E. Helou, M. Ibrahim, Queue-aware scheduling in full duplex ofdma wireless networks with imperfect channel state information, in: European Wireless 2018; 24th European Wireless Conference, 2018, pp. 1–7.
[27] X. Guo, G. Trimponias, X. Wang, Z. Chen, Y. Geng, X. Liu, Learning-based joint configuration for cellular networks, IEEE Internet Things J. 5 (6) (2018) 4283–4295.
[28] F. Tang, Y. Zhou, N. Kato, Deep reinforcement learning for dynamic uplink/downlink resource allocation in high mobility 5g hetnet, IEEE J. Sel. Areas Commun. (2020).
[29] X. Meng, H. Inaltekin, B. Krongold, Deep reinforcement learning-based power control in full-duplex cognitive radio networks, in: 2018 IEEE Global Communications Conference (GLOBECOM), IEEE, 2018, pp. 1–7.
[30] Y. Zhou, F. Tang, Y. Kawamoto, N. Kato, Reinforcement learning-based radio resource control in 5g vehicular network, IEEE Wirel. Commun. Lett. 9 (5) (2019) 611–614.
[31] C. Zhang, P. Patras, H. Haddadi, Deep learning in mobile and wireless networking: A survey, IEEE Commun. Surv. Tutor. (2019).
[32] H. Fawaz, K. Khawam, S. Lahoud, M. El Helou, A game theoretic framework for power allocation in full-duplex wireless networks, IEEE Access 7 (2019) 174013–174027.
[33] O. Osterb, Scheduling and capacity estimation in lte, 2011, pp. 63–70.
[34] K. Park, W. Willinger, Self-Similar Network Traffic and Performance Evaluation, 2000.
[35] S. Sesia, M. Baker, I. Toufik, LTE-the UMTS Long Term Evolution: From Theory to Practice, John Wiley & Sons, 2011.
[36] R.S. Sutton, A.G. Barto, Reinforcement Learning: An Introduction, 2011.
[37] M.N. Katehakis, A.F. Veinott Jr., The multi-armed bandit problem: decomposition and computation, Math. Oper. Res. 12 (2) (1987) 262–268.
[38] P.E. Mogensen, J. Wigard, Cost action 231: Digital mobile radio towards future generation system, final report, in: Section 5.2: On Antenna and Frequency Diversity in Gsm. Section 5.3: Capacity Study of Frequency Hopping Gsm Network, 1999.
[39] K. Pearson, The problem of the random walk, Nature 72 (1867) (1905) 342.
[40] R. McGill, J.W. Tukey, W.A. Larsen, Variations of box plots, Amer. Statist. 32 (1) (1978) 12–16.

Hassan Fawaz received a diploma in Telecommunication Engineering from the Lebanese University in 2015, and a master's degree in Telecom Networks and Security from Saint Joseph University of Beirut in 2016. He received a Ph.D in Telecommunication Engineering in 2019 from Saint Joseph University of Beirut with his research focusing on resource allocation in full-duplex wireless networks. In 2020, he completed a PostDoc at UVSQ, Paris-Saclay, France, with focus on IoT LoRaWAN Networks. Currently, he is a research engineer at Telecom SudParis-IMT, France.

Melhem El Helou received the engineer's degree and master's degree in Telecommunications and Networking Engineering from Ecole Supérieure d'Ingénieurs de Beyrouth (ESIB), Saint Joseph University of Beirut, Beirut, Lebanon, in 2009 and 2010, respectively and the Ph.D. degree in Computer and Telecommunications Engineering from IRISA Research Institute, University of Rennes 1, France and Saint Joseph University of Beirut, in 2014. He joined ESIB in September 2013 where he is currently an Assistant Professor (fr: Maître de conférences). His teaching and research interests include wireless networking, cellular technologies, Internet of Things, and quality of service.

Samer Lahoud received the Ph.D. degree in communication networks from IMT Atlantique, Rennes, in 2006. After his Ph.D. degree, he spent one year at Nokia Bell Labs Europe. From 2007 to 2016, he was with the University of Rennes 1 and with IRISA Rennes as an Associate Professor. He is currently an Associate Professor with the Saint Joseph University of Beirut, where he lectures computer networking courses with the Faculty of Engineering, Ecole Supérieure d'Ingénieurs de Beyrouth (ESIB). His research activities focus on routing and resource allocation algorithms for wired and wireless communication networks.

Kinda Khawam got her engineering degree from Ecole Superieure des Ingenieurs de Beyrouth (ESIB) in 2002, the Master's degree in computer networks from Telecom ParisTech (ENST), Paris, France, in 2003, and the Ph.D. from the same school in 2006. She was a post doctoral fellow researcher in France Telecom, Issy-Les-Moulineau, France in 2007. Actually, she is an associate professor and researcher at the University of Versailles in France. Her research interests include radio resource management, modeling and performance evaluation of mobile networks.