Andrea Ortiz, Arash Asadi, Max Engelhardt, Anja Klein, Matthias Hollick, “CBMoS: Combinatorial Bandit Learning for Mode Selection and Resource Allocation in D2D Systems”, in *IEEE Journal on Selected Areas in Communications.*, October 2019.

CBMoS: Combinatorial Bandit Learning for Mode Selection and Resource Allocation in D2D Systems

Andrea Ortiz†, Arash Asadi*, Max Engelhardt*, Anja Klein†, Matthias Hollick*
*Secure Mobile Networking lab (SEEMOO), Technische Universität Darmstadt
{firstname.lastname}@seemoo.tu-darmstadt.de
†Communication Engineering Lab, Technische Universität Darmstadt
{firstname.lastname}@nt.tu-darmstadt.de

***Abstract*—The complexity of the *mode selection and resource allocation (MS&RA)* problem has hampered the commercialization progress of Device-to-Device (D2D) communication in 5G networks. Furthermore, the combinatorial nature of MS&RA has forced the majority of existing proposals to focus on constrained scenarios or offline solutions to contain the size of the problem. Given the real-time constraints in actual deployments, a reduction in computational complexity is necessary. Adaptability is another key requirement for mobile networks that are exposed to constant changes such as channel quality fluctuations and mobility. In this article, we propose an online learning technique (i.e., CBMoS) which leverages combinatorial multi-armed bandits (CMAB) to tackle the combinatorial nature of MS&RA. Furthermore, our two-stage CMAB design results in a tight model, which eliminates the theoretically feasible but practicality invalid options from the solution space. We prototype the first SDR-based D2D testbed to verify the performance of CBMoS under real-world conditions. The simulations confirm that the fast learning speed of CBMoS leads to outperforming the benchmark schemes by up to 132%. In experiments, CBMoS exhibits even higher performance (up to 142%) than in the simulations. This stems from the adaptability/fast learning speed of CBMoS in presence of high channel dynamics which cannot be captured via statistical channel models used in the simulators.**

I. Introduction

Device-to-Device (D2D) communication emerged as a *disruptive* paradigm allowing direct communication between user equipments (UEs) without traversing the eNodeB (eNB), enabling public-safety, IoT, industry 4.0, and V2X scenarios [1]. After a decade of research and standardization effort, D2D is still not ready for commercial implementation. To date, 3GPP has only defined the overall architecture and basic functionalities (e.g., user/service discovery) [2]. The slow progress is endowed to the complexity of the *mode selection and resource allocation (MS&RA)* problem and the intertwined interference environment of D2D communication modes [3, 4]. The optimization-based solutions for MS&RA does not always achieve the time-constraints expected in a real-world scenario. Furthermore, they need to be recomputed every time the network conditions changes (e.g., channel variation).

A. Background and related work

By definition, there are three available D2D modes, namely, *inband*, *outband*, and *legacy*, see Fig. 1. In inband mode, D2D communication occurs over the LTE-A wireless interface of the UE in the licensed spectrum. Outband mode consists in communicating over the WiFi interface in unlicensed ISM spectrum [1]. Legacy mode refers to the legacy communication through the eNB, as it happens in today's networks, which is the least desirable D2D mode since relaying the transmission through the eNB consumes both uplink and downlink resources. Inband and outband modes are more resource-efficient since the D2D UEs can directly communicate without a relay in between. However, interference management is difficult in inband mode as D2D communication over licensed band can negatively impact other UEs in case of frequency reuse. Outband mode, on the other hand, is susceptible to interference from other devices and technologies operating in ISM band. As result, the selection of a suitable D2D mode is highly dependent on the channel dynamics of the surrounding environment. Note that MS&RA goes beyond a simple selection among these three modes. MS&RA entails deciding the communicating mode, the resources to be used within the selected mode, and the possibility of reusing the selected resources (by accounting for interference). Given the complexity of this decision, it comes as no surprise that the optimal MS&RA is an NP-Hard problem [5, 6].

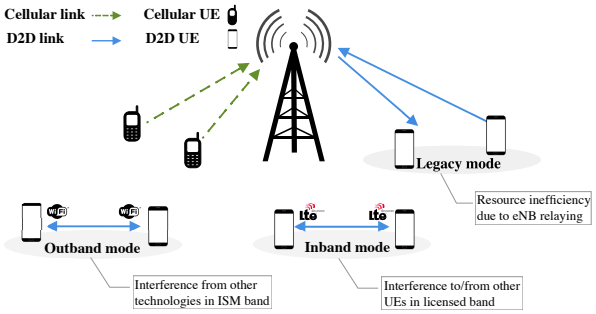


Fig. 1: An overview of different D2D modes.

To date, MS&RA problem has been addressed via mathematical tools such as classic optimization [6–10], game theory [11–13], and graph theory [14–16]. The seminal works in [17, 18] propose heuristics on mode selection between inband and legacy modes aiming at reducing interference from cellular UEs to D2D UEs in static scenarios. In [10], the authors formulate the D2D resource allocation as a non-convex optimization problem which is then solved via approximation. The authors of [7–9] follow similar approach in formulating the D2D mode selection as a (non)-linear optimization problem, which is solved through heuristics due to NP-Hardness. Furthermore, the authors of [7] derive a throughput-optimal solution for a network with only one D2D user in static scenarios. In [11] and [13], the authors use

matching theory and stackelberg games, respectively, to solve MS&RA problem between inband and legacy modes. In [12], the inband D2D resource allocation problem is formulated as a non-convex bargaining game theory framework, which is solved via iterative methods. The authors of [14] show that regular graph-theory approaches [15, 16] are too complex for solving D2D resource allocation, and propose an inband resource allocation algorithm using hypergraphs.

The majority of the above works only focus on D2D resource allocation, which in essence reduces the complexity of the MS&RA problem to scheduling D2D UEs within one mode. Among those that focus on MS&RA, important practical aspects such as accounting for network dynamics [3, 6, 12, 13, 17, 18] and computational complexity of the problem [15, 16, 19] are not addressed.

B. Motivation and challenge

A practical D2D MS&RA should *maintain low complexity while accounting for network dynamics (e.g., channel variation and mobility) in all D2D modes*. However, accounting for network dynamics in all three D2D modes intensifies the complexity of an already convoluted problem. The limitation of the state of the art in terms of computational complexity stems in the combinatorial nature of mode selection problems. Furthermore, the problem formulation with interference consideration is often non-linear which is even harder to solve. Although *reduced and simplified* versions of this problem can be solved via numerical methods, its complexity remains an issue, and its scalability is a concern. Hence, the majority of the state of the art aims at (semi) static scenarios in inband mode [20].

Leveraging machine learning techniques for MS&RA can facilitate higher adaptability to network dynamics. However, similar to aforementioned mathematical tools, the combinatorial nature of mode selection impacts the performance of learning algorithms by increasing the action space. In particular, learning decelerates because many feasible actions should be explored before converging to the best solution. The major challenge is devising a practical learning methods while accounting for the following:

- **Adapting to network dynamics.** Existing solutions often solve the MS&RA problem for a snapshot of the system and cannot cope with network dynamics [3, 6, 9, 12, 13, 17, 18]. Thus, any change in the system demands re-solving the problem without benefiting from the prior knowledge [4].
- **Considering all D2D modes.** The majority of the literature only focuses on the inband mode [7–10, 14–18]. However, 3GPP includes the support for outband from Rel.12 to encourage interworking of WLAN and LTE-A [1, 3]. Extending support to all D2D modes adds new dimensions to the problem.
- **Minimizing computational complexity.** Many proposals demonstrate impressive results, even with bounded optimality. Unfortunately, such solutions are often offline algorithms which do not meet the millisecond-scale scheduling requirement of cellular networks [9, 15, 16].

C. An overview of CBMoS

In this paper, we propose CBMoS, an online learning technique for *joint MS&RA* in D2D networks. CBMoS is designed based on a novel two-stage combinatorial multi-armed bandit (CMAB) approach. From an algorithmic point of view, the separation into stages allow us to reduce the action space and increase the learning speed of the algorithm. The first stage considers the resource allocation for cellular UEs. By exploiting the fact that resources cannot be shared among cellular UEs, the first stage breaks the combinatorial nature of the resource allocation problem and it is able to find solutions that aim at maximizing the throughput. The second stage handles the MS&RA for D2D UEs. In this case, the dimensionality of the problem is addressed by learning the best resource allocation for every D2D pair and evaluating the effect of the combined solution on the overall network throughput. Furthermore, to explore the possible actions efficiently, the two stages of CBMoS leverage a D2D-specific action selection policy based on a gradient-ascent approach.

From a networking point of view, CBMoS will simply replace the existing scheduling algorithm at the eNB. Located at the eNB, CBMoS has access to the channel quality indicator (CQI) of the UEs and can therefore determine the achieved performances in every scheduling slot. Using this information, CBMoS updates its learning parameters, i.e., its estimates on the expected performance of the possible MS&RA solutions, and uses them to select the MS&RA of each UE in each decision interval. The selected MS&RA solution is broadcast via a DCI channel, as it is done in LTE-A. Note that this decision contains the RB allocation for the cellular UEs, RB allocation and mode selection for legacy, inband and outband D2D UEs. After each decision interval, the UEs report back the CQI and the outcome of the decision, i.e., the achieved performance is obtained. Using this feedback, CBMoS updates its learning parameters in order to select the next MS&RA solution.

D. Our contributions

The following summarizes the contributions of this paper:

- We model the joint MS&RA problem as a two-stage CMAB, see Section II. Our novel formulation prunes the infeasible solutions by design via decoupling the decisions for cellular and D2D UEs. Such decoupling enables the reduction of the action space significantly.
- We propose CBMoS which provides fast convergence by leveraging a tailored action selection policy, see Section III. Specifically, the proposed policy uses the gradient-ascent approach to learn preferences for each possible MS&RA solution. These preferences are based on the throughput achieved in the network when the corresponding MS&RA was selected and ensure a more efficient exploration of the action space.

- We develop the first SDR-based testbed that is capable of full inband and outband D2D communication, see Section IV. Our modular design of the testbed allows for future extensions and implementation of new algorithms with minimal efforts. We have made the code base for inband D2D publicly available[1].
- The evaluation shows that CBMoS is the fastest to adapt and the first to identify best strategies under variant user density, network size, and channel occupancy. Furthermore, CBMoS achieves up to 132% and 142% throughput gain in simulations and experiments, respectively. The superiority of CBMoS stems from the aforementioned two-stage design and the proposed preference-based action selection policy. The former eliminates infeasible actions without incurring additional overhead while the latter exploits past experience in the calculation of the preferences. That is, the performance achieved by each selected MS&RA solution is used in the update of all the preferences, thus reducing the number of iterations required to identify efficient solutions.

II. System model

We aim to solve the joint MS&RA problem for D2D and cellular UEs. In this section, we first describe the network model and D2D modes in detail. Next, we elaborate on the learning model and the underlying technical details which tie our problem to a CMAB formulation. Finally, we formally define the proposed two-stage CMAB.

A. Network model

We consider a multi-cell LTE-A network with 20 MHz bandwidth in which downlink and uplink operate on separates channels (i.e., FDD[2]) under a Rayleigh fading channel. We assume a micro-cell deployment for urban areas in which the inter-site distance is 200 m [21]. The eNB provides coverage to N UEs, from which N_{c} are cellular UEs and N_{d} are D2D UEs. Every UE is equipped with one LTE-A interface and one WiFi interface. The traffic from the cellular UEs goes outside the network, whereas a D2D UE communicates with another UE within the cell (see Fig. 1). We consider three modes, namely, legacy, inband, and outband. Each mode contains a limited set of resources. The number of legacy and inband resource blocks (RBs) is denoted by R_{c} and R_{i}, respectively. The number of outband resources R_{o} is the number of available WiFi channels.

Legacy mode. In this mode, the D2D transmits the data via the uplink to the eNB, which then relays the packet to the D2D receiver via a downlink. Although the D2D pair in legacy mode communicate via the eNB (similar to the cellular UEs), there is a key difference between D2D and cellular users. The resource used by the cellular users cannot be re-used within the same cell, whereas the resources used by D2D users can be reused.

[1]http://sine.ni.com/cs/app/doc/p/id/cs-17689

[2]Since CBMoS only looks at the achieved performance on individual RBs, we can apply the same solution to TDD.

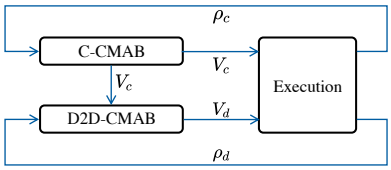


Fig. 2: Schematic of the two-stage CMAB problem

Inband mode. In this mode, the D2D pair communicates directly via their LTE-A interface over inband RBs. Unlike legacy mode, D2D pairs are allowed to reuse the resources within the cell in inband mode. Hence, D2D pairs can potentially interfere with each other.

Outband mode. In this mode, the D2D pair communicates directly over one of the available WiFi channels. In outband mode, the D2D pairs use contention-based medium access (i.e., CSMA), and thus their throughput can be impacted by other existing users/services operating in the same band. Note that the outband link is established following WiFi Direct association procedure [22] and 3GPP specification on WLAN-LTE integration as described in [23, 24]. After link activation, the D2D pair can report the channel statistics over LTE interface to the eNB.

B. Learning model

In essence, the MS&RA problem consists in choosing a resource within one of the available modes for every UE in the cell. Thus, the MS&RA problem matches very well to classic multi-armed bandit (MAB) formulations, where a decision-making agent must repeatedly choose one of several MS&RA solutions (i.e., arms). However, the large number of possible solutions, which depend on the amount of UEs in the network, makes a classic MAB formulation infeasible for a practical MS&RA scheme. That is, MAB requires each arm to be tried several times to learn the optimal solution. For instance, the number of solutions for a small network with 16 UEs is 2^{91} when the three modes are considered. In such a case, trying each possible MS&RA solution takes 7.8×10^{16} years assuming each action is verified within 1 ms. Therefore, to overcome this limitation, we model the D2D MS&RA as a two-stage CMAB problem, as depicted in Fig. 2.

The classic MAB formulation ignores the inter-dependency of the individual solutions (e.g., the impact of UE1's MS&RA on UE2) while CMAB accounts for this inter-dependency by looking at the composite outcome of these solutions. Given the high inter-dependency of D2D MS&RA decisions, CMAB facilitates faster and smarter learning than the classic MAB. Our proposed two-stage design stems from the fact that MS&RA for cellular UEs only entails resource allocation since they do not use D2D modes. Our formulation considers this by decoupling MS&RA of cellular UEs from D2D UEs. As a result, we reduce the solution space and improve the learning speed. Specifically, we perform resource allocation for the cellular UEs in the first stage (i.e., C-CMAB). Next, the output of C-CMAB is fed into the second stage (i.e., D2D-CMAB) to perform MS&RA for the D2D UEs. After the execution of every MS&RA decision, the eNB records the

TABLE I: Table of notations

Symbol	Notation	Symbol	Notation
i	Index for variable X	j	Index for the super-arms
k	Index for the value (arm) of variable X_i	t	Index for time interval
v_i	Selected value (arm) for variable X_i	$v_{i,k}$	k^{th} possible value (arm) of variable X_i
$H(v_{i,k})$	Preference of selecting arm $v_{i,k}$ for X_i	I	Total number of variables X_i
J	Total number of super-arms	J_{c}	Number of super-arms in C-CMAB
J_{d}	Number of super-arms in D2D-CMAB	K_i	Total number of values (arms) variable X_i can take
L	Validation function	N	Total number of UEs
N_{c}	Total number of cellular UEs	N_{d}	Total number of D2D UEs
R_{c}	Number of legacy resource blocks	R_{i}	Number of inband resource blocks
R_{o}	Number of outband resource blocks	V_j	j^{th} super-arm
X_i	Variable for which a decision has to be made	$\mathcal{V}$	Set of super-arms
$\mathcal{V}_i$	Set of values (arms) for variable X_i	α	Learning rate
ϵ	Probability of selecting the actions via Local MABs	μ_j	Expected reward of super-arm V_j
ρ_j	Reward function of super-arm V_j	$\bar{\rho}_j$	Estimated average reward of super-arm V_j
ϱ_i	Reward function of arm v_i	$\bar{\varrho}_i$	Estimated average reward of arm v_i

observed performance (i.e., throughputs) for updating learning parameters.

C. Problem formulation: Two-stage CMAB

A CMAB is usually determined by a set of variables, a set of arms and super-arms, a reward function and a validation function, as described in the following:

- The set of I variables $X = \{X_1, ..., X_I\}$ are the elements in the problem for which a decision has to be made. For C-CMAB, the variables X_i correspond to the available R_{c} legacy resource blocks and for D2D-CMAB the variables X_i are the $N_{\text{d}}/2$ D2D-pairs. For each variable X_i there are K_i different options (i.e., arms) and the value of X_i depends on the arm that is being selected. For example, if X_i represents a D2D UE, the arms correspond to the available resources and modes. The collection of the possible values X_i can take in each time interval is given by the set $\mathcal{V}_i = \{v_{i,1}, ..., v_{i,K_i}\}$, with $|\mathcal{V}_i| = K_i$.
- The collection of the values taken by all the I variables is termed a super-arm and it is given by the vector $V = (v_1, ..., v_I)$. Furthermore, the set containing all the S possible super-arms is defined as $\mathcal{V} = \{V \in \mathcal{V}_1 \times ... \times \mathcal{V}_I\}$. Following our previous example, when X_i represents D2D UEs, a super-arm V corresponds to the collection of the resources and modes selected by each D2D UE. In summary, an arm is associated to a single variable while a super-arm is the collection of the arms selected by all the variables.
- The reward function $\rho_j : V_j \rightarrow \mathbb{R}$ is a random variable with an unknown distribution and an expected value μ_j. It is determined for each super-arm $V_j \in \mathcal{V}$ and represents how beneficial is the selection of super-arm V_j in a given time interval.
- The validation function L determines whether a super-arm $V_j \in \mathcal{V}$ is a valid solution or not. Although validation functions introduce overhead to the learning algorithm, they are needed when the collection of the values taken by the variables X_i leads to practically infeasible solutions. In our case, the feasibility of the super-arms is determined by the possibility of sharing resources among the cellular and D2D UEs. As a consequence, we propose a tight model in which the unfeasible solutions are excluded and the use of a validation function is successfully avoided.

We have summarized all the variables and notation in Table I. As depicted in Fig. 2, our two-stage CMAB problem decouples the MS&RA of the D2D and cellular UEs by considering two different CMAB problems, namely C-CMAB and D2D-CMAB.

C-CMAB. As described, the C-CMAB performs resource allocation for the cellular UEs. Each variable X_i, $i = 1, ..., R_{\text{c}}$, corresponds to one available resource in legacy mode. Since each resource can only be allocated to one cellular UE at a time, the set $\mathcal{V}_i$ of arms available for each X_i is formed by the set of cellular UEs. Consequently, the number of possible arms is the same for all the resources X_i and it is given by the total number of cellular UEs, i.e., $K_i = N_{\text{c}}$. Furthermore, in C-CMAB, each super-arm is formed by the values taken by each of the R_{c} variables. In other words, the collection of the scheduled UEs (i.e., the UEs which received a resource). As every variable can take one out of N_{c} values, the total number of possible super-arms $|\mathcal{V}| = J_{\text{c}}$ can be calculated as the product of the possible values each resource can take which can be expressed as:

$$J_{\text{c}} = (N_{\text{c}})^{R_{\text{c}}}. \tag{1}$$

The reward obtained when selecting a super-arm V_j, $j = 1, ..., J_{\text{c}}$ is the total throughput achieved by cellular UEs[3]. As mentioned, our two-stage design removes the practically infeasible solutions (e.g., a cellular UE using inband resources). As a result, all the J_{c} super-arms are valid solutions for the MS&RA and the definition of a validation function L is unnecessary. The output of the C-CMAB, i.e., the resource allocation for the cellular UEs, is fed into the second stage to determine which cellular resources are available for the D2D UEs.

D2D-CMAB. The second stage performs MS&RA for the D2D UEs. Unlike cellular UEs, the reuse of resources within a mode is allowed for the D2D UEs. Consequently, the D2D-CMAB considers $I = \frac{N_{\text{d}}}{2}$ variables X_i, $i = 1, .., \frac{N_{\text{d}}}{2}$ where each variable X_i represents a D2D pair. The set $\mathcal{V}_i$ of arms available for each D2D pair is formed by the collection of possible mode selection solutions. The size of $\mathcal{V}_i$ represents the number of

[3]By means of a different reward function, other metrics such as proportional fairness or delay can be considered. The reward should reflect the suitability of the action taken with respect to the goal, e.g., for proportional fairness higher rewards should be obtained when the distribution of resources is more uniform.

possible solutions which can be expressed as:

$$|\mathcal{V}_i| = K_i = R_\mathrm{o} + \sum_{l=1}^{R_\mathrm{i}} \binom{R_\mathrm{i}}{l} + \sum_{l=1}^{R_\mathrm{c}} \binom{R_\mathrm{c}}{l} + 1. \quad (2)$$

The first two terms in Eq. (2) are the possible MS&RA options for outband and inband resources, respectively. The third term depends on the number of cellular resources. This represents the scenario in which D2D pairs operate in the legacy mode, and 1 accounts for the case when no resources are allocated to the D2D pair. The MS&RA of all D2D UEs forms a super-arm in this stage. The total number of super-arms $|\mathcal{V}_\mathrm{d}| = J_\mathrm{d}$ can be calculated as:

$$J_\mathrm{d} = \prod_{i=1}^{N_\mathrm{d}/2} K_i = \left(R_\mathrm{o} + \sum_{l=1}^{R_\mathrm{i}} \binom{R_\mathrm{i}}{l} + \sum_{l=1}^{R_\mathrm{c}} \binom{R_\mathrm{c}}{l} + 1 \right)^{N_\mathrm{d}/2}. \quad (3)$$

Similar to the C-CMAB stage, the reward obtained when selecting super-arm V_j is the total throughput achieved by D2D UEs. Moreover, the definition of the function L is unnecessary since all the super-arms are valid MS&RA solutions.

III. CBMoS Algorithm

We propose CBMoS, an online learning algorithm which is designed to solve the MS&RA in D2D systems. The online nature of CBMoS makes it effective in face of wireless channel dynamics. In its core, CBMoS exploits the Naive Sampling (NS) strategy [25]. The NS strategy breaks the CMAB problem into several smaller MAB problems. This separation is based on the assumption that for any CMAB, the reward function ρ_j of super-arm V_j can be approximated as the summation of the individual reward functions $\varrho(v_i)$ of its arms as follows:

$$\rho_j \approx \sum_{i=1}^{I} \varrho(v_i),\ v_i \in V_j. \quad (4)$$

In our two-stage problem, the assumption in (4) is fulfilled with equality by both, C-CMAB and D2D-CMAB. This is due to the fact that the total throughput is the summation of the achieved throughput of D2D and cellular UEs. Moreover, this throughput per UE already includes the effect of the interference caused by the other transmitting UEs as well as the interference caused by other cells using the same frequency bands. Following the formulation in Section II-C, CBMoS solves the MS&RA in two stages which are elaborated in the following. The pseudo-code of CBMoS is shown in Algorithm 1.

A. First stage: C-CMAB

As shown in Fig. 3a, NS is used to separate C-CMAB into R_c local and one super-MAB problems. Each local-MAB problem is associated with a particular variable X_i, $i = 1, ..., R_\mathrm{c}$ which corresponds to a cellular resource. At every decision interval $t = 1, ..., T$, each local-MAB selects a cellular UE from its corresponding set $\mathcal{V}_i$. Note that the local-MAB problems do not have a combinatorial nature because they only consider one variable at a time. As explained in Section II, the collection of the arms selected by all the local-MAB problems forms a super-arm V_j. For C-CMAB, the super-arm contains the scheduled cellular UEs. The task of the super-MAB is to select one super-arm from the ones that have been already observed.

Algorithm 1 Two-stage CMAB algorithm

```
1: initialize α, α_ϱ, ϵ, H and
2: for every t = 1, ..., T do
3:     generate random number ζ                                  ▷ C-CMAB
4:     if ζ ≥ ϵ(t) then                                  ▷ Exploit from super-MAB
5:         select the super-arm with higher expected reward
6:     else                                              ▷ Explore from local-MABs
7:         for each local-MAB do
8:             calculate the preferences of each arm             ▷ Eq. (7)
9:             calculate the probabilities for each arm          ▷ Eq. (8)
10:            select the arm with the highest preference
11:        end for
12:        collect the selected arms to form the super-arm
13:        if a new super-arm is encountered then
14:            add it to the set of available super-arms in super-MAB
15:        end if
16:    end if
17:    determine available resources in legacy mode
18:    generate random number ζ                                  ▷ D2D-CMAB
19:    if ζ ≥ ϵ(t) then                                  ▷ Exploit from super-MAB
20:        select the super-arm with higher expected reward
21:    else                                              ▷ Explore from local-MABs
22:        for each local-MAB do
23:            calculate the preferences of each arm             ▷ Eq. (7)
24:            calculate the probabilities for each arm          ▷ Eq. (8)
25:            select the arm with the highest preference
26:        end for
27:        if a new super-arm is encountered then
28:            add it to the set of available super-arms in super-MAB
29:        end if
30:    end if
31:    observe the achieved throughput
32:    update baseline in C-CMAB and D2D-CMAB                    ▷ Eq. (6)
33:    update super-arms' expected rewards                       ▷ Eq. (5)
34: end for
```

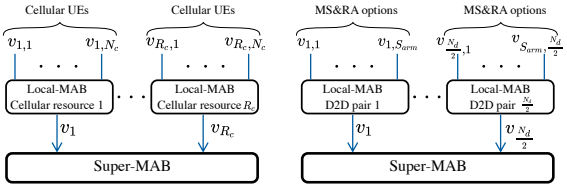
As a result, the number of available super-arms grows every time a new V_j is encountered. In summary, the local-MAB problems are used to *explore* and discover new super-arms, and the super-MAB is used to *exploit* the already known super-arms by selecting the best available super-arm.

To balance the trade-off between exploration and exploitation, we consider the ϵ-greedy policy. At every decision interval t, the super-MAB selects a super-arm with probability $1 - \epsilon$ (line 4) or the local-MAB selects a super-arm with probability ϵ (line 6). The super-MAB follows a greedy policy for the selection of the super-arms (line 5), i.e., the super-MAB always selects the *best* super-arm from the set of available super-arms. Furthermore, the quality of each super-arm V_j, $V_j \in \mathcal{V}$ is measured in terms of the estimate of its average reward $\bar{\rho_j}$, which is calculated via the following:

$$\bar{\rho}_{j,t+1} = \bar{\rho}_{j,t} + \alpha \left(\rho_{j,t} - \bar{\rho}_{j,t} \right), \quad (5)$$

where $0 \leq \alpha < 1$ is a fraction that affects the learning rate and $\rho_{j,t}$ is calculated using (4). Note that using $\alpha = 1/t$ in (5) results in the sample-average method.

For the arm selection in the local-MAB, we propose a customized preference-based policy, in which we obtain the preference of allocating resource X_i, $i = 1, ..., R_\mathrm{c}$ to the k^{th}, $k = 1, ..., N_\mathrm{c}$ cellular UE (line 8). This policy increases the convergence rate of CBMoS because it promotes the selection of arms whose achieved throughput is higher than the average throughput per resource experienced up to decision interval t



(a) C-CMAB (b) D2D-CMAB

Fig. 3: Schematic of the NS strategy for C-CMAB and D2D-CMAB

($\bar{\varrho}_{i,t}$). We denote the difference between the achieved throughput $\varrho_{i,t}$ when v_i is selected and the average throughput per resource $\bar{\varrho}_{k,t}$, $k = 1, ..., I$, by:

$$\delta_{k,t} = \varrho_{i,t} - \bar{\varrho}_{k,t}. \tag{6}$$

CBMoS learns the preferences based on the idea of gradient ascent [26]. With (6), the preferences of all the arms in a local-MAB are updated every time t as follows:

$$H_{t+1}(v_{i,k}) = \begin{cases} H_t(v_{i,k}) + \alpha\delta_{k,t}\left(1 - \mathbb{P}(v_{i,k})\right) & \text{if } X_i = v_{i,k} \\ H_t(v_{i,k}) - \alpha\delta_{k,t}(\mathbb{P}(v_{i,k})), & \text{if } X_i \neq v_{i,k} \end{cases}, \tag{7}$$

where $\mathbb{P}(v_{i,k})$ is the probability of selecting arm $v_{i,k}$ and it can be calculated using a soft-max distribution as:

$$\mathbb{P}\left(X_i = v_{i,k}\right) = \frac{e^{H_t(v_{i,k})}}{\sum_{l=1}^{K_i} e^{H_t(v_{i,l})}}. \tag{8}$$

In every interval, CBMoS checks whether the selected super-arm in local-MAB has already been encountered (line 13). When a new super-arm is detected, the set of available super-arms in super-MAB is updated (line 14). After the selection of the super-arm, CBMoS determines the number of available resources in legacy mode (line 17). This number determines the possible MS&RA for the D2D pairs in the next stage (i.e., D2D-CMAB).

B. Second stage: D2D-CMAB

Although the techniques used in D2D-CMAB are similar to that of C-CMAB, they vary in the definition of the variables. For brevity, we focus on these differences. Fig. 3b shows that the NS divides D2D-CMAB into $\frac{N_d}{2}$ local and one super-MAB problems. In this stage, each local-MAB represents a D2D pair. At every decision interval, each local-MAB selects an arm which is the MS&RA for its corresponding D2D pair. The number of available arms S_{arm} is computed via (2). The collection of the MS&RA for each D2D pair forms a super-arm. As mentioned, the task of the super-MAB is to select one super-arm from the ones that have been already observed.

We use ϵ-greedy to decide whether to select a super-arm via the super-MAB or the local-MAB problems (line 19). Similar to C-CMAB, the selection of D2D-CMAB consists in the use of greedy policy for the super-MAB and the preference-based policy for the local-MAB problems. In D2D-CMAB, $H_t(v_{i,k})$ is the preference of allocating the resources in arm $v_{i,k}$, $k = 1, ..., S_{\text{arm}}$ to the i^{th}, $i = 1, ..., \frac{N_d}{2}$, D2D pair. In each iteration, CBMoS checks whether the selected super-arm in D2D-CMAB has already been encountered (line 27).

Finally, after the selection of the super-arms in C-CMAB and D2D-CMAB is performed, the MS&RA solution is applied and the achieved throughput is observed. This achieved throughput is used by CBMoS to update the averages and expected reward estimates in C-CMAB and D2D-CMAB (lines 31-33).

A common metric to evaluate the performance of the strategies used to solve CMAB problems is the regret. The regret is defined as the expected loss caused by the fact that the optimal super-arm is not always selected [27]. The regret at decision interval t is calculated as:

$$\eta_t = \sum_{j=1}^{J} (\mu^* - \mu_j)\mathbb{E}[m_{j,t}] = \sum_{j=1}^{J} \Delta_j \mathbb{E}[m_{j,t}] \tag{9}$$

where $\Delta_j = \mu^* - \mu_j$ and $m_{j,t}$ is the number of times super-arm V_j has been selected up to decision interval t. The analysis of the regret can be found in the following subsection.

C. Regret Analysis

In this section, we first derive the regret of C-CMAB and D2D-CMAB. Next, we show the regret of CBMoS. For these derivations, the seminal work of [27] is used as a baseline.

Proposition 1. *For C-CMAB and D2D-CMAB, the probability of selecting a non-optimal super-arm V_j after t decision intervals, $t \geq a - b$, where $a, b \in \mathbb{R}^+$, $a \geq b$ are constant values, is bounded by*

$$\mathbb{P}\left(V_j\right) \leq \frac{a}{b+t} + 2a\log\left(\frac{e^{1/a}(b+t)}{a}\right) + \frac{4}{d^2}\left(\frac{a}{e^{1/a}(b+t)}\right)^{\frac{ad^2}{2}}, \tag{10}$$

where $d = \min_{j:\mu_j<\mu^*}\{\mu^* - \mu_j\}$ is the difference of between the expected reward of the optimal super-arm V^* and the best non-optimal super-arm V_j.

Proof. The probability of selecting super-arm V_j at any given decision interval t is given by

$$\mathbb{P}(V_j) \leq \epsilon_t \prod_{i=1}^{I} \mathbb{P}(X_i = v_i) + (1 - \epsilon_t)\mathbb{P}(\bar{\mu}_{j,t-1} \geq \bar{\mu}^*_{t-1}), \tag{11}$$

where $\bar{\mu}_{j,t}$ and $\bar{\mu}^*_t$ are the estimated average reward of V_j and V^*, respectively, at decision interval t. The first term in (11) corresponds to the probability of selecting V_j via the local MABs, while the second term is the probability of selecting V_j via the super-MAB. For the first term, we consider the worst case for the upper bound. This is, that a sub-optimal arm V_j will be selected when exploring. As a result,

$$\mathbb{P}(V_j) \leq \epsilon_t + (1 - \epsilon_t)\mathbb{P}(\bar{\mu}_{j,t-1} \geq \bar{\mu}^*_{t-1}). \tag{12}$$

To calculate the upper bound of the second term in (11), we exploit the fact that

$$\mathbb{P}(\bar{\mu}_{j,t} \geq \bar{\mu}^*_t) \leq \mathbb{P}\left(\bar{\mu}_{j,t} \geq \lambda\right) + \mathbb{P}\left(\bar{\mu}^*_t \leq \lambda\right), \tag{13}$$

where $\lambda = \frac{\mu^* + \mu_j}{2}$. Now, for each term on the right side of (13) we have that,

$$\mathbb{P}\left(\bar{\mu}_{j,t} \geq \lambda\right) = \sum_{l=1}^{t} \mathbb{P}\left(m_{j,t} = l \wedge \bar{\mu}_{j,t} \geq \lambda\right) \tag{14}$$

$$= \sum_{l=1}^{t} \mathbb{P}\left(m_{j,t} = l | \bar{\mu}_{j,t} \geq \lambda\right)\mathbb{P}\left(\bar{\mu}_{j,t} \geq \lambda\right). \tag{15}$$

Using Hoeffdings's inequality, the second term in (15) can be bounded as

$$\mathbb{P}\left(\bar{\mu}_{j,t} \geq \lambda\right) \leq e^{\frac{-\Delta_j^2 l}{2}}, \tag{16}$$

As a result,

$$\mathbb{P}\left(\bar{\mu}_{j,t} \geq \lambda\right) \leq \sum_{l=1}^{t} \mathbb{P}\left(m_{j,t} = l | \bar{\mu}_{j,t} \geq \lambda\right) e^{\frac{-\Delta_j^2 l}{2}}. \tag{17}$$

The right side of (17) can be written as

$$\sum_{l=1}^{t_0} \mathbb{P}\left(m_{j,t} = l | \bar{\mu}_{j,t} \geq \lambda\right) e^{\frac{-\Delta_j^2 l}{2}} + \sum_{l=t_0+1}^{t} \mathbb{P}\left(m_{j,t} = l | \bar{\mu}_{j,t} \geq \lambda\right) e^{\frac{-\Delta_j^2 l}{2}}. \tag{18}$$

After bounding some terms in (18) to one, we obtain

$$\sum_{l=1}^{t_0} \mathbb{P}\left(m_{j,t} = l | \bar{\mu}_{j,t} \geq \lambda\right) + \sum_{l=t_0+1}^{t} e^{\frac{-\Delta_j^2 l}{2}}. \tag{19}$$

Moreover, the exponential term in (19) can be bounded such that

$$\mathbb{P}\left(\bar{\mu}_{j,t} \geq \lambda\right) \leq \sum_{l=1}^{t_0} \mathbb{P}\left(m_{j,t} = l | \bar{\mu}_{j,t} \geq \lambda\right) + \frac{2}{\Delta_j^2} e^{\frac{-\Delta_j^2 t_0}{2}} \tag{20}$$

holds. Denoting $m_{j,t}^{\text{rnd}}$ as the number of times super-arm V_j has been randomly selected, (20) can be expressed as

$$\mathbb{P}\left(\bar{\mu}_{j,t} \geq \lambda\right) \leq t_0 \mathbb{P}\left(m_{j,t}^{\text{rnd}} \leq t_0\right) + \frac{2}{\Delta_j^2} e^{\frac{-\Delta_j^2 t_0}{2}}. \tag{21}$$

However, note that in C-CMAB the super-arm V_j can only be randomly selected one time, i.e., at the beginning of the algorithm when the preferences have not yet been calculated. As a result, $\mathbb{P}\left(m_{j,t}^{\text{rnd}} \leq t_0\right) = 1$ and

$$\mathbb{P}\left(\bar{\mu}_{j,t} \geq \lambda\right) \leq t_0 + \frac{2}{\Delta_j^2} e^{\frac{-\Delta_j^2 t_0}{2}}. \tag{22}$$

Now, let $t_0 = \sum_{l=1}^{t} \epsilon_l$. Moreover, as $\epsilon_t = \frac{a}{b+t}$ where $a, b \in \mathbb{R}^+$ and $a \geq b$, a lower bound for t_0 can be found by considering $t' = b - a$ as follows:

$$t_0 = \sum_{l=1}^{t'} \frac{a}{b+l} + \sum_{l=t'+1}^{t} l = 1 \frac{a}{b+l} \tag{23}$$

$$t_0 \geq 1 + a \log\left(\frac{b+t}{b+t'}\right) \tag{24}$$

$$t_0 \geq a \log\left(\frac{e^{1/a}(b+t)}{a}\right). \tag{25}$$

Thus, using (12), (13), (22) and (25), and writing $d = \min_{j:\mu_j < \mu^*}\{\mu^* - \mu_j\}$, the probability in (11) can be expressed as

$$\mathbb{P}\left(V_j\right) \leq \epsilon_t + \left(2t_0 + \frac{4}{\Delta_j^2} e^{\frac{-\Delta_j^2 t_0}{2}}\right) \tag{26}$$

$$\mathbb{P}\left(V_j\right) \leq \frac{a}{b+t} + 2a \log\left(\frac{e^{1/a}(b+t-1)}{a}\right) + \frac{4}{d^2}\left(\frac{a}{e^{1/a}(b+t-1)}\right)^{\frac{ad^2}{2}} \tag{27}$$

□

Theorem 1. *The regret of C-CMAB and D2D-CMAB is bounded by*

$$\eta_t \leq \sum_{j=1}^{J} \Delta_j \left(\frac{a}{b+t} + 2a \log\left(\frac{e^{1/a}(b+t-1)}{a}\right) + \frac{4}{d^2}\left(\frac{a}{e^{1/a}(b+t-1)}\right)^{\frac{ad^2}{2}}\right), \tag{28}$$

where $a, b \in \mathbb{R}^+$, $a \geq b$ are constant values and $d = \min_{j:\mu_j < \mu^*}\{\mu^* - \mu_j\}$ is the difference of between the expected reward of the optimal super-arm V^* and the best non-optimal super-arm V_j. The leading order of the regret is $O\left(2a \log\left(\frac{e^{1/a}(b+t-1)}{a}\right)\right)$.

Proof. From (9), the regret can be rewritten as

$$\eta_t = \sum_{j=1}^{J} \Delta_j \mathbb{P}(V_j). \tag{29}$$

The bound obtained by using Proposition 1:

$$\eta_t \leq \sum_{j=1}^{J} \Delta_j \left(\frac{a}{b+t} + 2a \log\left(\frac{e^{1/a}(b+t-1)}{a}\right) + \frac{4}{d^2}\left(\frac{a}{e^{1/a}(b+t-1)}\right)^{\frac{ad^2}{2}}\right), \tag{30}$$

in which the leading order is given by

$$O\left(2a \log\left(\frac{e^{1/a}(b+t-1)}{a}\right)\right). \tag{31}$$

□

Proposition 2. *For CBMoS, the probability of selecting a non-optimal combination of super-arms $V_{\text{c},j}$ and $V_{\text{d},j}$, for C-CMAB and D2D-CMAB, respectively after t decision intervals, $t \geq a - b$, where $a, b \in \mathbb{R}^+$, $a \geq b$ are constant values, is bounded by*

$$\mathbb{P}(V_{\text{c},j} \cup V_{\text{d},j}) \leq \mathfrak{P} \tag{32}$$

where $\mathfrak{P}$ is defined in (33), $d_{\text{c}} = \min_{j:\mu_{\text{c},j} < \mu_{\text{c}}^*}\{\mu_{\text{c}}^* - \mu_{\text{c},j}\}$ is the difference of between the expected reward of the optimal super-arm V_{c}^* and the best non-optimal super-arm $V_{\text{c},j}$ in C-CMAB and $d_{\text{d}} = \min_{j:\mu_{\text{d},j} < \mu_{\text{d}}^*}\{\mu_{\text{d}}^* - \mu_{\text{d},j}\}$ is the difference of between the expected reward of the optimal super-arm V_{d}^* and the best non-optimal super-arm $V_{\text{d},j}$ in D2D-CMAB.

Proof. For CBMoS, the probability of selecting a non-optimal combination of super-arms is

$$\mathbb{P}(V_{\text{c},j} \cup V_{\text{d},j}) = \mathbb{P}(V_{\text{c},j}) + \mathbb{P}(V_{\text{d},j}) + \mathbb{P}(V_{\text{c},j})\mathbb{P}(V_{\text{d},j}), \tag{34}$$

where $\mathbb{P}(V_{\text{c},j})$ and $\mathbb{P}(V_{\text{d},j})$ are the probabilities of selecting non-optimal super-arms in C-CMAB and D2D-CMAB, respectively. Using the result in Proposition 1 and after some algebraic manipulations, (34) can be bound as

$$\mathbb{P}(V_{\text{c},j} \cup V_{\text{d},j}) \leq \mathfrak{P}. \tag{35}$$

□

Theorem 2. *The regret of CBMoS is bounded by*

$$\eta_t \leq \sum_{j=1}^{J^{\text{CBMoS}}} \Delta_j \mathfrak{P}, \tag{36}$$

where $\mathfrak{P}$ is defined in (33), J^{CBMoS} is the total number of combined super-arms, $d_{\text{c}} = \min_{j:\mu_{\text{c},j} < \mu_{\text{c}}^*}\{\mu_{\text{c}}^* - \mu_{\text{c},j}\}$ is the difference of between the expected reward of the optimal super-arm V_{c}^* and the best non-optimal super-arm $V_{\text{c},j}$ in C-CMAB and $d_{\text{d}} = \min_{j:\mu_{\text{d},j} < \mu_{\text{d}}^*}\{\mu_{\text{d}}^* - \mu_{\text{d},j}\}$ is the difference of between the expected reward of the optimal super-arm V_{d}^* and the best non-optimal super-arm $V_{\text{d},j}$ in D2D-CMAB. The leading order of the regret is $O\left(4a^2 \log^2\left(\frac{e^{1/a}(b+t-1)}{a}\right)\right)$.

$$\mathfrak{P} = \frac{2a}{b+t} + 4a\log\left(\frac{e^{1/a}(b+t)}{a}\right) + \frac{4}{d_{\mathrm{c}}^2}\left(\frac{a}{e^{1/a}(b+t)}\right)^{\frac{ad_{\mathrm{c}}^2}{2}} + \frac{4}{d_{\mathrm{d}}^2}\left(\frac{a}{e^{1/a}(b+t)}\right)^{\frac{ad_{\mathrm{d}}^2}{2}} + \left[\left(\frac{a}{b+t} + 2a\log\left(\frac{e^{1/a}(b+t)}{a}\right) + \frac{4}{d_{\mathrm{c}}^2}\left(\frac{a}{e^{1/a}(b+t)}\right)^{\frac{ad_{\mathrm{c}}^2}{2}}\right)\left(\frac{a}{b+t} + 2a\log\left(\frac{e^{1/a}(b+t)}{a}\right) + \frac{4}{d_{\mathrm{d}}^2}\left(\frac{a}{e^{1/a}(b+t)}\right)^{\frac{ad_{\mathrm{d}}^2}{2}}\right)\right], \quad (33)$$

Proof. From (9), the regret can be rewritten as

$$\eta_t = \sum_{j=1}^{J^{\mathrm{CBMoS}}} \Delta_j \mathbb{P}(V_{\mathrm{c},j} \cup V_{\mathrm{d},j}). \quad (37)$$

The bound is obtained by using Proposition 2:

$$\eta_t \leq \sum_{j=1}^{J^{\mathrm{CBMoS}}} \Delta_j \mathfrak{P} \quad (38)$$

From there it is easy to see that the leading order is given by

$$O\left(4a^2\log^2\left(\frac{e^{1/a}(b+t-1)}{a}\right)\right). \quad (39)$$

□

D. Computational complexity evaluation

In this section, we evaluate the computational complexity of one iteration of our proposed CBMoS with respect to the size of the network, i.e., the number of UEs (N_{c} and N_{d}) and the number of available resources (R_{c}, R_{i} and R_{o}). We start our analysis with C-CMAB and then continue with D2D-CMAB.

From Algorithm 1, it can be observed that for C-CMAB, the selection of super-arms via the super-MAB (line 5) entails the selection of the super-arm with the maximum average estimated reward. This operation has a complexity that grows as $O(J_{\mathrm{c},t})$, where $J_{\mathrm{c},t} \leq J_{\mathrm{c}} = N_{\mathrm{c}}^{R_{\mathrm{c}}}$ is the number of super-arms in the super-MAB up to decision interval t. For the selection of super-arms via the local-MABs, three steps need to be followed. First, the preferences are calculated (line 8). As this operation is performed for each arm, its complexity grows as $O(N_{\mathrm{c}})$. Second, the probabilities for each arm are calculated (line 9). This step has a complexity that grows as $O(N_{\mathrm{c}})$ because it involves the calculation of N_{c} probabilities. Note that the calculation of the denominator in (8) is done only once and the result is stored in memory. Third, the arm with the highest preference is selected (line 10). As mentioned before, determining the maximum value out of N_{c} values has a complexity that grows as $O(N_{\mathrm{c}})$. Consequently, as these three steps need to be repeated R_{c} times (line 7), the complexity of the selection of the super-arm via the local-MABs grows as $O(R_{\mathrm{c}}N_{\mathrm{c}})$. Similarly, for D2D-CMAB, the selection of super-arms via the super-MAB (line 20) has a computational complexity that grows as $O(J_{\mathrm{d},t})$, where $J_{\mathrm{d},t} \leq J_{\mathrm{d}} = \left(R_{\mathrm{o}} + \sum_{l=1}^{R_{\mathrm{i}}}\binom{R_{\mathrm{i}}}{l} + \sum_{l=1}^{R_{\mathrm{c}}}\binom{R_{\mathrm{c}}}{l} + 1\right)^{N_{\mathrm{d}}/2}$ is the number of super-arms in the super-MAB up to decision interval t. Let us denote $R_{\mathrm{d}} = R_{\mathrm{o}} + \sum_{l=1}^{R_{\mathrm{i}}}\binom{R_{\mathrm{i}}}{l} + \sum_{l=1}^{R_{\mathrm{c}}}\binom{R_{\mathrm{c}}}{l} + 1$. Then we have that for the selection of super-arms via the local-MABs, the complexity grows as $O(R_{\mathrm{d}}N_{\mathrm{d}}/2)$. This is because each of the required operations (lines 23-25) have a complexity that grows linearly with R_{d} and each the operations needs to be repeated $N_{\mathrm{d}}/2$ times. To determine how does the computational complexity of CBMoS grow, we consider the most computationally demanding ways to select the super-arms in both, C-CMAB and D2D-CMAB, i.e., the selection of super-arm via the super-MABs. As a result, the complexity of CBMoS grows linearly with the number of super-arms as $O(J_{\mathrm{c},t} + J_{\mathrm{d},t})$.

Following a similar approach, we can obtain that the complexity of the benchmark approaches (ϵ-greedy and UCB) grows as $O(J_t)$ where $J_t \leq J = J_{\mathrm{c}}J_{\mathrm{d}}$. This means, the two-stage design of CBMoS allow us to achieve a complexity that grows linearly with the number of possible super-arms, instead of the quadratic growth of the benchmark approaches.

IV. SDR Design and Implementation

In this section, we elaborate on the overall architecture and hardware/software components of the testbed. In addition, we briefly explain the process of prototyping inband and outband D2D communication. Finally, we provide details on the integration of CBMoS in our testbed.

Our SDR hardware is comprised of NI 2954R USRP RIO devices[4]. These devices are equipped with Xilinx Kintex-7 FPGAs and multiple RF daughterboards. The USRPs are connected through a PCI-e interface to a real-time host, which is an off-the-shelf PC running NI Linux RT (NILRT), a Linux real-time operating system. We use LabVIEW Communications to program and control the platforms both in FPGA and NILRT. Specifically, the computation-intensive parts of our project (e.g., DSP blocks) are programmed in FPGA using LabVIEW's Clock-Driven Logic programming language. Real-time host code mostly comprises higher-layer control functionalities which require deterministic execution in μ-second scale (e.g., scheduling). Our testbed benefits from the LabVIEW Application Frameworks (AFWs), which are reference designs of LTE and WiFi physical layers [5,6].

Overview of the eNB. The eNB is comprised of a real-time host and a USRP, as depicted in Fig. 4. The main responsibilities of the real-time host are processing the feedback and signaling messages received from the UEs. This information is used as input for our MS&RA algorithms. Furthermore, the real-time host generates the transport blocks to be transmitted via the physical layer. These transport blocks are sent to the USRP for (de-)coding, (de-)modulation and RF transmission and reception. Moreover, physical layer operations such as channel equalization are performed at the USRP.

Overview of the UE. As shown in Fig. 5, the UE is comprised of a real-time host and two USRPs: one for emulation

[4]http://www.ni.com/de-de/support/model.usrp-2954.html
[5]http://www.ni.com/white-paper/53286/en/
[6]http://www.ni.com/white-paper/53279/en/

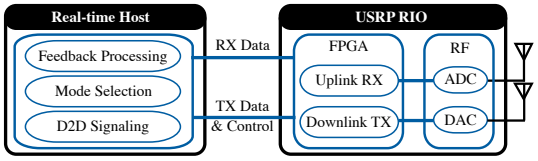


Fig. 4: High-level FPGA architecture of our D2D-enabled eNB

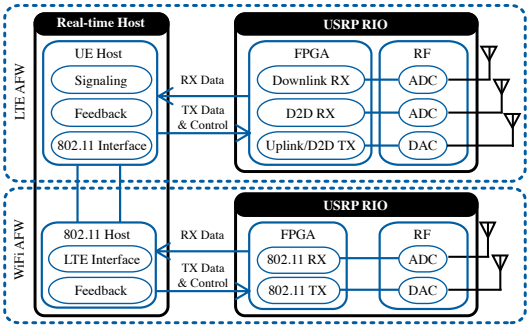


Fig. 5: High-level FPGA architecture of our D2D-enabled UE

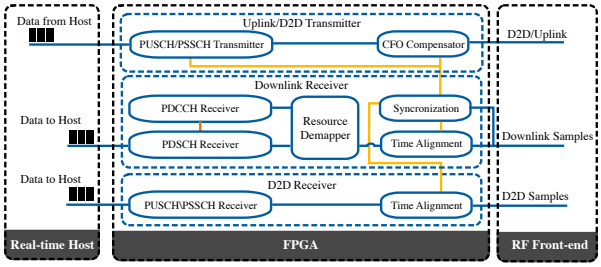


Fig. 6: Overview of inband D2D implementation at the UE

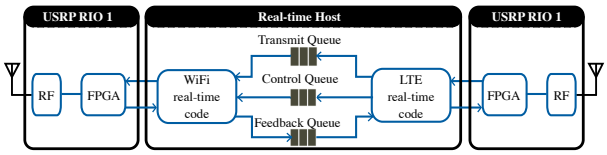


Fig. 7: Communication between LTE and WiFi AFW

of the LTE and the other for the WiFi interface, respectively. The FPGA logic required for both physical layers exceeds the available fabric on a Kintex-7 FPGA, hence the use of two USRPs. The real-time host at the UE runs both LTE and WiFi AFWs and facilitates *communication between the two*. Similar to the eNB, the USRPs at the UE handle baseband and RF processing, such as carrier frequency offset (CFO) compensation. In addition to the physical layer, the USRP running the WiFi AFW code implements a CSMA MAC layer.

Although the LTE and WiFi reference design implementations are instrumental to our testbed, numerous extensions are required to enable inband and outband support. These extensions include adding support for Multi-UE, enabling OFDMA, inband channel, and outband channel. For brevity, we provide a high-level description of the setup and only describe the most significant changes for inband and outband.

A. Inband D2D communication links

According to the 3GPP specification, inband D2D users communicate over the *sidelink channel* [28]. In our implementation, we leverage the FPGA logic of uplink transmitter (at the UE) and receiver (at the eNB) to implement the sidelink transceiver. Given that sidelink uses part of the uplink resources, the UE either uses the uplink or sidelink channel in every given subframe. We have to make the following modifications to the transmission time interval (TTI) management block to add this feature. Fig. 6 shows the high-level overview of our inband D2D implementation at the UE.

The Host-FPGA interface is extended to contain additional parameters for the D2D link including a ten-slot *TX subframe configurations* array, which specifies for each subframe whether it is to be used for uplink, sidelink, or no transmission at all (e.g., to listen to an inbound D2D transmission). Depending on the subframe index, the TTI handling block outputs the respective configuration for the current subframe.

As mentioned, we leverage the existing uplink logic to implement sidelink. However, this task is not as easy as duplicating the eNB's uplink receiver logic in the UE, since we need to account for synchronization. The eNB determines system timing and carrier frequencies in LTE. The UEs synchronize themselves with the eNB and estimate the CFO by processing the primary synchronization signal (PSS). The primary sidelink synchronization signal (PSSS) is a separate synchronization signal which is used for D2D communications according to [28]. We deviate from the standard in our implementation by using the PSS signal for synchronization. The LTE resource grid implementation in mostly hard-coded and adding PSSS signal changes this grid which in turn sets off a chain of modifications. Hence, we refrained from altering the resource grid due to timing constraints. It should be noted that this will not impact the accuracy of our results since PSSS is important, particularly in out-of-coverage scenarios. These scenarios are out of the scope of this paper.

Fig. 6 depicts the FPGA architecture of our D2D-enabled UEs. The CFO and timing offset information obtained from processing the PSS in the *Synchronization* block in the downlink receiver chain are shared with the uplink transmitter chain and the sidelink transceiver chain. The AFW already includes the CFO correction and time alignment for the uplink transmitter chain. We port this feature to the sidelink receiver.

B. Outband D2D communication links

We use the WiFi AFW to implement outband D2D links. The architecture of outband links is shown in Fig. 7. We use real-time queues to communicate between the LTE and WiFi AFWs on host-level. For implementing the network-assisted WLAN control feature specified by the standard, we use three real-time queues: (i) *Control Queue* which provides the MAC addresses (source and destination) for the outband transmission; (ii) *Transmit Queue* which supplies the data to be sent over the outband link; and (iii) *Feedback Queue* which reports the performance of the outband link. The latter is then reported to the eNB via the LTE uplink. The host portion of the LTE AFW receives configuration packets from the eNB and configures its WiFi interface accordingly via these queues.

TABLE II: Parameters used in the evaluation

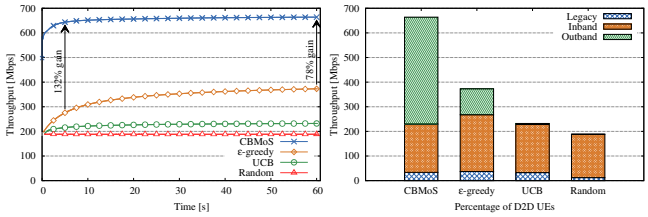
Parameter	Value
Cellular	
Cellular uplink bandwidth	20 MHz
Cell radius	200 m
TX power cellular UE	24 dBm
Thermal noise power	-174 dBm/Hz
Noise figure eNB, UEs	5, 7
Fading, shadowing, pathloss	Reyleigh, 4 dB, UMa [29]
Number of resource block pools	12
Outband	
WiFi bandwidth	22 MHz
WiFi effective range	150 m
WiFi TX power	20 dBm
D2D	
Underlay max bandwidth	20 MHz
Number of resource block pools	12
D2D maximum distance	30 m

C. Integration of CBMoS

The MS&RA occurs at the eNB, hence, CBMoS algorithm is deployed within the real-time host at the eNB. At the end of each decision interval, the UEs report the achieved throughput to the eNB over the uplink channel. This information will be available to CBMoS. Knowing the last decision, CBMoS can map the achieved throughputs to the previous MS&RA decision and calculate the rewards for the past actions. As for implementation of CBMoS, we leverage the mathscript node of LabVIEW which allows executing Matlab code within LabVIEW. In addition to reducing development time, this approach eliminates the performance difference due to efficiency of the code because we use the same implementation of CBMoS in the simulations and experiments. Note that integrating CBMoS in a real-deployment does not require changes to the architecture or signaling procedure of existing cellular networks since it will only replace the existing scheduling algorithm. The key difference is that the D2D UEs requires to send feedback for both licensed and unlicensed spectrum, whereas cellular users only send feedback for the licensed spectrum.

V. Evaluation

In this section, we evaluate and benchmark CBMoS via numerical simulations and experiments. Our simulation environment (e.g., cell size and wireless propagation) is created in accordance with 3GPP technical specification (see Sections 6.2 and 7.4 in [21] and Table II). We assume a multi-cell scenario with frequency re-use factor 1 in which the cell under evaluation can be interfered by its first-tier neighbors (6 cells) and second-tier (12 cells) neighboring cells [30]. Note that this is a worst-case scenario which maximizes the effect of intercell interference. The duration of each simulation is 60 s, and it is repeated for 500 times. The simulations demonstrate the performance of CBMoS in large networks (e.g., 100 UEs per cell) which cannot be studied using our testbed. Although our testbed consists of 20 USRP nodes, the maximum number of UEs in our experiments is limited to one eNB and eight UEs; each UE requires two USRPs. In addition, we use a couple of USRPs to create interference over outband channels. Unless otherwise specified, the occupancy of outband channel is 50%, and the number of UEs is 100 (50 cellular and 50 D2D UEs). All users move at pedestrian speed following a random direction mobility model [31]. The mobility not only results in instantaneous channel fluctuations but also changes the statistics behind these fluctuations, e.g., as the distance between the eNB and the UE increases, the probability of having low channel quality grows. As mentioned, no prior work has used learning algorithms to solve the joint MS&RA problem. Nevertheless, we adapt two well-known learning policies (i.e., ϵ-greedy and UCB) to the MS&RA problem as a benchmark to our proposal. For fair comparison, we adapt both strategies to a CMAB instead of the original MAB problem formulation.



(a) Throughput over time. (b) Throughput per mode.

Fig. 8: The learning speed and achieved throughput per mode.

- **ϵ-greedy.** In this approach, the ϵ-greedy policy is used for the selection of arms in the global and local MAB problems, i.e., for each MAB, there is a probability $(1-\epsilon_{\text{MAB}})$ of selecting the best arm and a probability ϵ_{MAB} of randomly selecting one of the available arms. Note that we have tailored this strategy such that only valid super-arms and arms can be selected (i.e., avoids infeasible solutions).
- **UCB.** This approach is similar to ϵ-greedy, but the UCB strategy in [27] is used in each local MAB for the selection of the arms.
- **Random.** At each decision interval, the algorithm selects randomly one of the valid MS&RA solution. Each solution is chosen with the same probability.

A. Learning speed

Fig. 8a illustrates the aggregate system throughput over time. We observe that CBMoS has much higher learning speed in comparison to the benchmark schemes. Specifically, CBMoS achieves up to 643 Mbps after 5 seconds and 662 Mbps after 60 seconds which is 132% and 78% higher than the most performant benchmark, respectively. Fig. 8b sheds light on the contribution of each mode in the aggregate system throughputs after 60 seconds. We can see that the least throughput is achieved in legacy mode because reusing resource blocks is not allowed in this mode. This figure illustrates the strategy difference among CBMoS and the benchmark schemes. In particular, we observe that CBMoS better utilizes the available resource in outband while the rest of the algorithms have more focus on the inband and cellular resources. Outband resources are not well-explored by ϵ-greedy and UCB because outband constitutes a smaller number of arms than inband and cellular, hence the lower exploration probability.

The large performance difference between CBMoS and the benchmark schemes stems from: (i) the two-stage CMAB

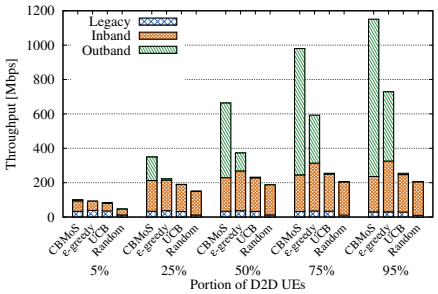


Fig. 9: Number of D2D UEs vs. throughput.

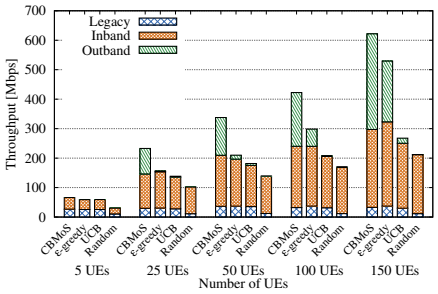


Fig. 10: Network size vs. throughput.

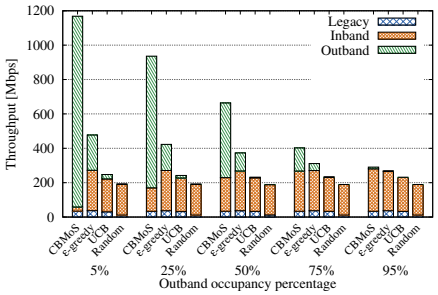


Fig. 11: Outband occupancy vs. throughput.

design which facilitates the formulation of learning problems with a smaller solution space and (ii) the preference-based action selection policy which enables more efficient exploration/exploitation of the MS&RA problem. Specifically, for this scenario, the number of possible solutions for CBMoS is at most $J_c + J_d = 50^{25} + (8192)^{25}$, while for the other schemes is at most $Jc J_d = (409600)^{25}$, i.e., approximately 2.98×10^{42} times higher.

Note that theoretically ϵ-greedy and UCB can achieve the same performance as CBMoS if they are given infinite time. Nevertheless, practical MS&RA demands for *fast and efficient solutions* while ϵ-greedy and UCB suffers from slow learning speed due to the prolonged exploration/exploitation phases even when they are embedded in a CMAB formulation. This is particularly noticeable for UCB because it forces continuous exploration until all actions are visited at least once. For example, in this scenario, each D2D pair has to try all the possible 8192 solutions at least once. Given the number of possible actions in the MS&RA, such exploration results in a slow learning curve.

B. Impact of number of D2D UEs

Here, we evaluate the impact of the percentage of D2D UEs (i.e., $5\%, 25\%, 50\%, 75\%, 95\%$) in the network of 100 UEs on the system throughput. Fig. 9 shows that the throughput significantly increases as the percentage of D2D UEs in the network grows. This observation tallies with the results depicted in Fig. 8b. Note that the reuse of resource is prohibited for cellular UEs but allowed for the D2D UEs in inband and outband mode. Hence, *the opportunity for reusing resources grows with the number of D2D UEs*. This result demonstrates the importance of D2D communication in enhancing cellular capacity.

C. Impact of network size

In the previous scenario, we observed that D2D UEs play an important role in the achieved system throughput. Next we study the impact of the number of UEs (i.e., network size) on the aggregate throughput. Fig. 10 shows that the aggregate throughput increases with the network size for all algorithms. As the network size increases, so does the number of D2D UEs (50% of the UEs are D2D UEs). As a result, as shown in Fig. 9, this leads to higher aggregate system throughput.

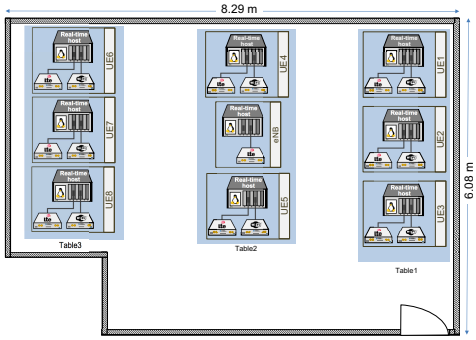


Fig. 12: The layout of the room and distribution of SDRs.

D. Impact of channel occupancy in outband

In this scenario, we analyze the impact of outband channel occupancy on the system throughput (see Fig. 11). Since outband channels are shared with all *other services* running on ISM frequencies, it is important to observe the system behavior in different traffic conditions. Here we evaluate a scenario in which the other services occupied $5\%, 25\%, 50\%, 75\%, 95\%$ of the outband channel capacity. The use of inband mode increases as the occupancy of the outband channel increases. We also observe that the performance gap among different algorithms reduces as the occupancy of outband channels increases. The reason behind the similarity of performances between different algorithms is two fold: firstly, outband resources, i.e., WiFi channels, constitute a large portion of the total resources, thus removing them pushes the system to a state in which there are not many resources to share among UEs and consequently, leaves little room for optimization. Secondly, the strength of CBMoS lies within its fast convergence and its ability to find the best solution in scenarios with very large action space. Removing outband significantly reduces the action space, which is one of the main challenges for the benchmarks. As a consequence, the gap in the performance between CBMoS and the benchmarks is reduced compared to the previous cases. Nevertheless, CBMoS still outperforms the benchmark schemes by at least 9%.

E. Experimental evaluation

To verify the practicality of CBMoS, we repeat the simulated scenario in our SDR-based testbed. The experiments are performed in a 50.4m^2 room. Our testbed is comprised of one eNB, which is placed in the center, and eight UEs which are distributed over three tables in the room, see Fig. 12.

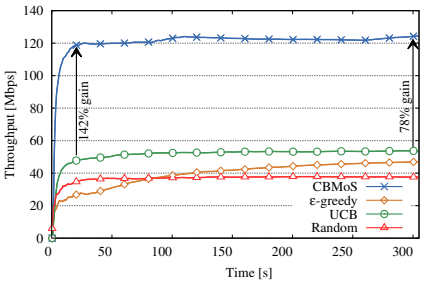


Fig. 13: The learning speed over time.

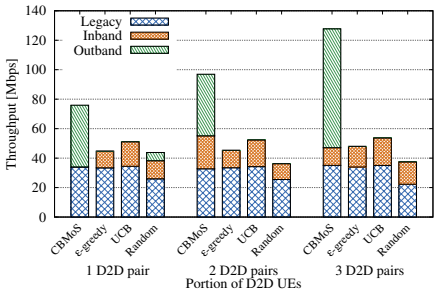


Fig. 14: Number of D2D pairs vs. throughput.

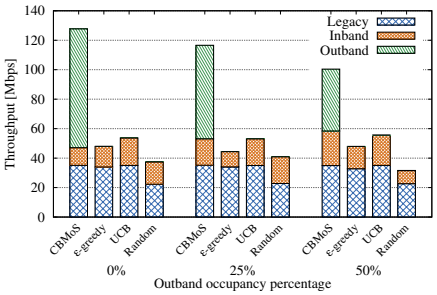


Fig. 15: Outband occupancy vs. throughput.

Fig. 13 demonstrates the aggregate system throughput achieved over 300 seconds for the different algorithms. In this scenario, there are three D2D pairs and two cellular UEs. We observe that CBMoS maintains its fast learning and high performance in experiments. Interestingly, CBMoS shows even higher performance in the experiments compared to the simulations. In experiments, we face unpredictable channel effects which cannot be captured by simulation models. While CBMoS's dynamic nature and its fast learning capabilities allows for quick adaption, the other benchmark schemes fail to cope with these changes.

Fig. 14 shows the impact of the number of D2D UEs in our experiments. As observed in the simulations, the performance of CBMoS increases with the number of D2D UEs due to the reuse of resources. Note that for one D2D pair, CBMoS learns to allocate one outband channel to the D2D pair because it provides, on average, a higher throughput than the inband resources. As mentioned before, the performance of ϵ-greedy and UCB is reduced compared to CBMoS due to their slow learning speed, as depicted in Fig. 13.

The effect of outband channel occupancy is presented in Fig. 15. It can be seen that the contribution of the outband channels to the aggregate throughput of CBMoS decreases with the increase of the occupancy, as shown in the simulation. Also, the usage of inband resources in CBMoS increases to compensate for this fact. The performance of the benchmarks is not significantly affected because, due to their slow learning speed, they do not exploit the available outband resources.

VI. Conclusions

In this article, we model the joint MS&RA problem as a two-stage combinatorial multi-armed bandit problem. Next, we propose CBMoS which leverages a D2D-specific exploration/exploitation policy. We benchmark CBMoS against popular learning approaches via extensive simulations in presence of network dynamics (e.g., fading, interference and mobility). To the best of our knowledge, there is no prior work proposing a practical solution for such dynamic networks. The simulation confirms that CBMoS increases the aggregate system throughput up to 2 folds. To verify the practicality of our proposal, we develop the first SDR-based testbed capable of inband and outband D2D communication. Our experimental evaluation confirms the high performance that is observed in the simulation.

In the course of this research, we modeled and solved the MS&RA via reinforcement learning and multi-armed bandits. We do not present these results due to space constraints. Nevertheless, both approaches resulted in very slow learning speed. To this aim, we adapt the benchmark algorithms to a combinatorial model to provide a fair comparison with CBMoS. The correct tuning of the learning parameters, e.g., ϵ and α, impacts the performance of the learning algorithms, hence they should be tuned in accordance with the intensity of the network dynamics. As a future work, we intend to derive these parameters on the fly from the reported performance values and their fluctuations.

VII. Acknowledgements

This work has been performed in the context of the DFG Collaborative Research Center (CRC) 1053 MAKI - subprojects A3, B3, and C1. This project has been partially funded by the LOEWE initiative (Hesse, Germany) within the NICER project. The authors would like to thank National Instruments, Dresden, in particular, Markus Unger, Clemens Felber, and Vincent Kotzsch for providing support in the testbed development.

References

[1] F. Jameel, Z. Hamid, F. Jabeen, S. Zeadally, and M. A. Javed, "A survey of device-to-device communications: Research issues and challenges," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 3, pp. 2133–2168, 2018.

[2] G. Araniti, A. Raschellà, A. Orsino, L. Militano, and M. Condoluci, "Device-to-device communications over 5G systems: Standardization, challenges and open issues," *5G Mobile Communications*, pp. 337–360, 2017.

[3] S. Andreev, D. Moltchanov, O. Galinina, A. Pyattaev, A. Ometov, and Y. Koucheryavy, "Network-assisted device-to-device connectivity: Contemporary vision and open challenges," in *Proceedings of EW*, 2015.

[4] F. S. Shaikh and R. Wismüller, "Routing in multi-hop cellular device-to-device (D2D) networks: A survey," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 2622–2657, 2018.

[5] T. D. Hoang, L. B. Le, and T. Le-Ngoc, "Resource allocation for D2D communications under proportional fairness," in *Proceedings of IEEE GLOBECOM*, 2014, pp. 1259–1264.

[6] A. Asadi, V. Mancuso, and P. Jacko, "Floating band D2D: Exploring and exploiting the potentials of adaptive D2D-enabled networks," in *Proceedings of IEEE WoWMoM*, 2015.

[7] S. Bulusu, N. B. Mehta, and S. Kalyanasundaram, "Rate adaptation, scheduling, and mode selection in D2D systems with partial channel knowledge," *IEEE TWC*, vol. 17, no. 2, pp. 1053–1065, 2018.

[8] G. Fodor, "Performance comparison of practical resource allocation schemes for device-to-device communications," *Wireless Communications and Mobile Computing*, 2018.

[9] A. Asadi, V. Mancuso, and R. Gupta, "DORE: An experimental framework to enable outband D2D relay in cellular networks," *IEEE/ACM Transactions on Networking*, vol. 25, no. 5, pp. 2930–2943, 2017.

[10] A. Ramezani-Kebrya, M. Dong, B. Liang, G. Boudreau, and S. H. Seyedmehdi, "Joint power optimization for device-to-device communication in

cellular networks with interference control," *IEEE TWC*, vol. 16, no. 8, pp. 5131–5146, 2017.
[11] S. M. A. Kazmi, N. H. Tran, W. Saad, Z. Han, T. M. Ho, T. Z. Oo, and C. S. Hong, "Mode selection and resource allocation in device-to-device communications: A matching game approach," *IEEE TMC*, vol. 16, no. 11, pp. 3126–3141, 2017.
[12] R. Wang, D. Cheng, G. Zhang, Y. Lu, J. Yang, L. Zhao, and K. Yang, "Joint relay selection and resource allocation in cooperative device-to-device communications," *International Journal of Electronics and Communications*, vol. 73, pp. 50–58, 2017.
[13] K. Zhu and E. Hossain, "Joint mode selection and spectrum partitioning for device-to-device communication: A dynamic stackelberg game," *IEEE TWC*, vol. 14, no. 3, pp. 1406–1420, 2015.
[14] H. Zhang, L. Song, and Z. Han, "Radio resource allocation for device-to-device underlay communication using hypergraph theory," *IEEE TWC*, vol. 15, no. 7, pp. 4852–4861, 2016.
[15] T. D. Hoang, L. B. Le, and T. Le-Ngoc, "Resource allocation for D2D communication underlaid cellular networks using graph-based approach," *IEEE TWC*, vol. 15, no. 10, pp. 7099–7113, 2016.
[16] R. Zhang, X. Cheng, L. Yang, and B. Jiao, "Interference-aware graph based resource sharing for device-to-device communications underlaying cellular networks," in *Proceedings of IEEE WCNC*, 2013.
[17] P. Janis, V. Koivunen, C. Ribeiro, J. Korhonen, K. Doppler, and K. Hugl, "Interference-aware resource allocation for device-to-device radio underlaying cellular networks," in *Proceedings of IEEE VTC*, 2009.
[18] H. Min, J. Lee, S. Park, and D. Hong, "Capacity enhancement using an interference limited area for device-to-device uplink underlaying cellular networks," *IEEE TWC*, vol. 10, no. 12, pp. 3995–4000, 2011.
[19] A. Asadi, P. Jacko, and V. Mancuso, "Modeling multi-mode D2D communications in LTE," *Proceedings of the workshop on mathematical performance modeling and analysis*, 2014.
[20] S. T. Shah, S. F. Hasan, B.-C. Seet, P. H. J. Chong, and M. Y. Chung, "Device-to-device communications: A contemporary survey," *Wireless Personal Communications*, vol. 98, no. 1, pp. 1247–1284, 2018.
[21] 3GPP, "Technical specification group radio access network: Study on channel model for frequencies from 0.5 to 100 GHz," TR 38.901, 2017.
[22] A. Asadi and V. Mancuso, "WiFi direct and LTE D2D in action," in *2013 IFIP Wireless Days (WD)*. IEEE, 2013, pp. 1–8.
[23] 3GPP, "Evolved universal terrestrial radio access (E-UTRA); medium access control (MAC) protocol specification," TS 36.321, 2018, v15.0.0.
[24] ——, "Evolved universal terrestrial radio access (E-UTRA); radio resource control (RRC); protocol specification," TS 36.331, 2018, v15.0.1.
[25] S. Ontañón, "The combinatorial multi-armed bandit problem and its application to real-time strategy games," in *Proceedings of AIIDE* , 2013.
[26] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT Press, 1998.
[27] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, no. 2-3, pp. 235–256, 2002.
[28] 3GPP, "Technical specification group radio access network; evolved universal terrestrial radio access (E-UTRA); physical channels and modulation (release 15)," TS 36.211, 2017, v15.0.0.
[29] ITU, "Guidelines for evaluation of radio interface technologies for IMT-2020," Tech. Rep., 2017.
[30] F. Z. Kaddour, E. Vivier, M. Pischella, and P. Martins, "A new method for inter-cell interference estimation in uplink SC-FDMA networks," in *Proceedings of IEEE VTC*, 2012.
[31] T. Camp, J. Boleng, and V. Davies, "A survey of mobility models for ad hoc network research," *Wireless communications and mobile computing*, vol. 2, no. 5, pp. 483–502, 2002.

Andrea Ortiz (S'14) received the B.S. degree in electronic engineering from Universidad del Norte, Barranquilla, Colombia, in 2008. In 2013 she obtained the M.S. degree in Information and Communication Engineering from Technische Universität Darmstadt, Darmstadt, Germany. Currently she is pursuing the Ph.D. degree at the Communications Engineering Lab, Technische Universität Darmstadt, Germany. Her research interests include reinforcement learning for wireless communications, signal processing for wireless communications and energy harvesting communications.

Dr. Arash Asadi received his Masters and Ph.D. degrees from IMDEA Networks Institute in 2012 and 2016, respectively. He is currently appointed as an Athena Young Investigator at Technische Universität Darmstadt. His research is focused on wireless communications both in sub-6Ghz and millimeter-wave bands as well as their applications in emerging areas such as vehicular communication and industry 4.0. He is a recipient of several awards including outstanding PhD and master thesis awards from UC3M. His papers on D2D communication have appeared in IEEE COMSOC best reading topics on D2D communication and in IEEE COMSOC Tech Focus.

Max Engelhardt holds a master degree in IT Security from Technische Universität Darmstadt. He currently works as a Software Development Engineer at Vector Informatik GmbH. His research interests include network security, 5G cellular networking, Device-to-Device (D2D) communication techniques and SDR prototyping.

Prof. Dr.-Ing. Anja Klein (M'96) received the Dr.-Ing. (Ph.D.) degree in electrical engineering from the University of Kaiserslautern, Germany, in 1996. In 1996, she joined Siemens AG, Mobile Networks Division, Munich and Berlin. She was active in the standardization of third generation mobile radio in ETSI and in 3GPP, e.g., leading the 3GPP RAN1 TDD group. She was director of a development and a systems engineering department. In 2004, she joined the Technische UniversitÃd't Darmstadt, Germany, as full professor, heading the Communications Engineering Lab. Her main research interests are in mobile radio, including interference management, cross-layer design, relaying and multi-hop, computation offloading, smart caching, and energy harvesting. Dr. Klein has authored over 280 peer-reviewed papers, has contributed to 12 books, and is inventor of more than 45 patents. In 1999, she was named the Inventor of the Year by Siemens AG.

Prof. Dr.-Ing. Matthias Hollick is currently heading the Secure Mobile Networking Lab in the Computer Science Department of Technische Universität Darmstadt, Germany. After receiving the Ph.D. degree from TU Darmstadt in 2004, he has been researching and teaching at TU Darmstadt, Universidad Carlos III de Madrid, and the University of Illinois at Urbana Champaign. His research focus is on resilient, secure, privacy-preserving, and quality-of-service-aware communication for mobile and wireless systems and networks.