# Efficient Mode Selection and Vehicle Pairing for Underlay V2X Networks

Zana Limani Fazliu*, Jeta Dobruna*, Hëna Maloku Berzati*, Carla Fabiana Chiasserini†, Francesco Malandrino‡
* University of Prishtina, Kosovo † Politecnico di Torino, Italy ‡ CNR-IEIIT, Italy

***Abstract*—Device-to-device communications (D2D) have been been a key driver in supporting vehicle-to-everything (V2X) cellular communication for vehicular networks. Underlay D2D communications in which devices reuse cellular spectrum in an opportunistic manner, face several challenges such as efficient mode selection and link stability in the presence of interference. In this paper we address the problem of mode selection and vehicle pairing for a V2X network underlaying a cellular multi-tier network. To solve the problem dynamically and independently from the network, while maximizing the spectral efficiency of the radio resources, we apply a reinforcement learning approach. The scheme we propose is fully decentralized and relies on a simplified yet efficient state space. The performance of the solution is evaluated through the simulation of a two-tier network operating in an urban environment, with realistic modeling of vehicular mobility. The performance of our approach is compared to two other benchmark approaches in terms of spectral efficiency and average user data rate. Our simulation results show that our approach can improve the spectral efficiency and average data rate for 90% of the vehicles in the network, while also ensuring around 12% of improvement in terms of overall network spectral efficiency.**

***Index Terms*—V2X, D2D mode selection, vehicle paring, reinforcement learning, spectral efficiency**

## I. Introduction

With the continuous increase of connected devices to the internet, data traffic has grown dramatically, imposing new requirements in terms of expected levels of Quality of Service (QoS) and delay tolerance. In response, operators have started deploying low power base stations (BSs) on top of the existing cells, to enhance coverage and capacity. The increasing densification of the network has been considered for some time now as one of the main enablers for next generation technologies.

In parallel, vehicles are increasingly emerging as some of the most data-hungry users of mobile networks. Facilitating vehicular to everything (V2X) communications, therefore is an important use case for 5G and beyond networks, especially with the advent of autonomous vehicles. V2X communications include all types of communications that can occur between vehicles themselves, i.e., V2V communications, as well as communications between the vehicles and infrastructure nodes (V2I), such as cellular network base stations (BS) and road side units (RSU).

In particular one of the main enablers for V2X communications in 5G networks is device-to-device (D2D) technology [1]. D2D communication refers to direct communication between devices, located near each other, without the intermediation of the network. These short-range communications have become one of the key technologies for boosting the performance of the current communication networks, gaining high research attention [2]. The use of D2D for V2X communications was introduced in 3GPP Release 14 and included in all subsequent standardization releases. D2D communications are also part of the Release 16 standardization efforts for 5G or *New Radio* (NR), in the specification for NR Sidelink, which is the physical interface used for D2D [3]. In the context of vehicular networks, D2D communications can be used to establish direct links between vehicles, or V2V links.

When D2D communications are deployed in an underlay fashion, the devices reuse the channel resources dedicated to cellular user equipment (UEs), hence enhancing spectrum efficiency. However, despite the obvious benefits, D2D users may cause interference to regular UE communications, introducing new challenges for network designers, such as interference management, resource and power allocation. Thus, to address these challenges and to take full advantage of D2D communication in current communication networks, resource, and power allocation algorithms must be carefully designed to guarantee QoS for both cellular and D2D users [4].

D2D links by their very nature are short-lived, especially between two moving devices, therefore it is important that the users are able to fall back to the network in case of link failure. Consequently, a device in D2D-supported networks may choose between two possible *communication modes*: *the D2D mode*, in which the device communicates with another device directly, or *the network mode*, in which the device is connected to a network BS. In the context of vehicular networks we will often refer to these modes as V2V and V2I, respectively. Furthermore, in case of D2D mode, another consideration is how to choose the best communicating peer. Both of these decisions can impact the level of QoS received as well as link stability.

In this work we tackle underlay V2X networks, in which reuse of cellular resources is allowed in an opportunistic fashion and coordination with the network is not required. In particular we address the problem of communication mode selection and propose a joint mode selection and vehicle pairing scheme using reinforcement learning. We consider the vehicle pairing problem in addition since we aim to achieve a fully network-independent solution, which would allow for dynamic reselection of D2D peers.

The scheme we propose is fully decentralized and relies on a simple model for the observation of the system state, without additional overhead. The scenario we consider is a cellular network composed of one macro cell and several pico cells,

ISBN: 978-3-903176-71-3

with support for D2D communications. Specifically, we focus on the use case of D2D for non-safety V2V communications, in an urban environment.

The rest of the paper is organized as follows. In Sec. II we present a review of relevant literature. In Sec. III we present the system model and provide a formulation of the mode selection and vehicle pairing problem. In Sec. IV, a reinforcement approach to tackle the problem is proposed, and finally in Sec. V we provide the results of the numerical simulations that were used to evaluate the performance of the proposed solution. Conclusions and future research directions are presented in Sec. VI.

## II. Related work

The problem of communication mode selection and resource allocation, especially in the context of V2X communications, has been an active area of research. Several works have addressed mode selection [5]–[7] and dynamic resource allocation strategies [8]–[10]. In [5], authors discuss mode switching strategies based on load information and signal strength, and analytically evaluate the impact of different strategies on performance indicators such as delay. In [6], the authors propose a deep reinforcement learning (DRL) algorithm to jointly address the communication mode selection as well as resource allocation. However both consider only a single cell cellular network. Authors in [7] also consider both mode selection and resource allocation for a vehicular network, however they focus on multicast transmissions between a cluster of vehicles, mainly for the purpose of sharing safety-critical information. Special attention has been paid to resource allocation solutions that do not rely on network aid. The authors in [8] show that opportunistic use of resources through dynamic scheduling performs significantly better for non-periodic types of traffic. Opportunistic use of resources without network coordination is addressed in [9] and [10], however like all of the above works, they too consider that V2V pairs are predetermined. Therefore none of the above mentioned works tackle the problem of efficient vehicle pairing.

Machine learning (ML) has recently emerged as a promising solution for 5G and 6G communication networks to optimize network resources and enhance system performance.

In [11] a user association scheme based on multi-agent reinforcement learning (RL) is proposed for maximizing the sum rate of mmWave network. The scheme is simulated in a two-tier network (composed of small and macro cells) considering all types of interference. On the other hand, authors in [12] proposed a decentralized user association technique based on multi-agent deep RL (DRL) to maximize the energy efficiency of ultra dense networks. However, none of these works consider D2D communication.

In [13] authors proposed an algorithm for the mode selection of vehicular users, using the DRL technique. The proposed algorithm aims to guarantee the high capacity of vehicle-to-infrastructure (V2I) links and ultra-reliability of D2D links under the QoS constraints. In this scenario only D2D users are considered, while interference from cellular users is not taken into account. In [14] and [15], authors adopted the RL technique to optimize resource allocation in a scenario considering both UEs and D2D users. In [14] a user association and resource allocation scheme is proposed to maximize the overall data rate of both cellular users and D2D pairs, using reinforcement learning. The scenario consists of only micro cells. In [15] the RL technique is also employed to maximize the network capacity in a D2D-enabled wireless network, through resource allocation, under the QoS constraints. This work too considers a scenario composed only of small cells.

Authors in [16] proposed an algorithm based on the DRL technique for improving resource allocation and power control for D2D users in a scenario composed of one-size cells. This algorithm considers QoS requirements and reduces the interference. In [17] the deep Q-network (DQN) is adopted to address the channel selection and power allocation issue in a D2D overlay communication network. This work aims to reduce the interference caused by sharing channels between D2D users.

**Novelty.** The literature review shows that ML techniques for future networks can deliver better results than traditional methods in terms of network performance. In this work, unlike similar other works that address the mode selection problem [13], [14], we model the mode selection and vehicle pairing as an extension to the user association problem. In particular, we do not consider the vehicle pairs that partake in D2D communications to be predetermined by the network, but rather as a variable that can be dynamically selected. We propose a solution that aims to improve vehicular network capacity while ensuring better overall spectral efficiency.

## III. System Model

In this work, we consider a two-tier network composed of a single macro cell and $L$ pico cells. The set of all cells is denoted with $\mathcal{S}$. User equipment (UEs) are distributed randomly within the coverage of the macro cell. Pico base stations are also randomly dropped within the coverage area of the macro cell. We denote the set of UEs as $\mathcal{C}$. Further, we consider a vehicular network composed of vehicular users (VUEs) placed randomly in two four-lane intersecting streets, moving with an average speed of 30km/h, as presented in Figure 1. The set of vehicular users is denoted with $\mathcal{V}$. The VUEs can operate as regular UEs or as VUEs but can use only one mode of communication at a time, therefore $\mathcal{V}$ is a subset of $\mathcal{C}$. We assume that a pair of VUEs $(v_t, v_r)$ can establish only one D2D link at a time. We assume that VUEs use uplink cellular resources for establishing D2D links as proposed by 3GPP [19], for non-safety related communications. We consider a frequency division duplex (FDD) system that uses $N_{UR}$ resource blocks (RB) for uplink communications, and a separate set of RBs, $N_{DR}$, for downlink communications.

In the context of NR, standards also recognize two modes for resource allocation procedure referred to as Mode 1 and Mode 2. In Mode 1, communication mode selection and resource allocation is completely coordinated by the network, including the selection of vehicle pairs that will communicate

ISBN: 978-3-903176-71-3

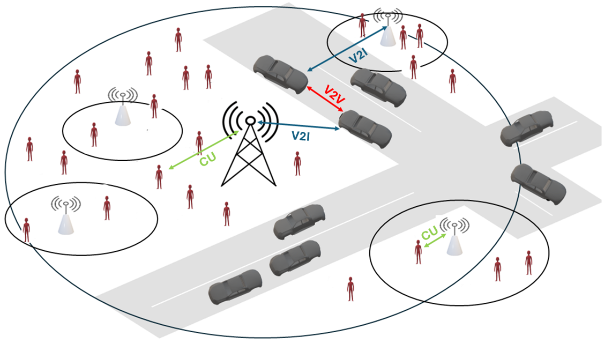

Fig. 1: System model. Vehicles in our scenario can communicate with a BS (V2I mode), and with other vehicles (V2V mode).

directly. Commonly, the network also allocates orthogonal radio resources to such D2D links, in order to minimize interference and degradation of its uplink communications. In Mode 2, by contrast, the resources of the cellular network are used opportunistically, and without explicit coordination with the cellular network, usually via a sensing procedure [18]. While many works consider that link establishment between predetermined vehicle pairs is supported by the network, we make no such assumptions in this work.

The UEs and VUEs transmit with a maximum power of $P_{tx}^{max}$, that is evenly spread out over the used RBs. The path loss model between the UEs and base stations,is calculated according to 3GPP models [20] for urban areas, depending on the cell size. For the channel between two users, either cellular or vehicular, we use the urban micro (UMi) model with a correction offset to account for the lower transmitter antenna height [21]. It should be noted that the expressions provided below are also indexed in the time domain, however to simplify the notation, the time index has been omitted. For a selected UE, $c$, which can also be a VUE in network mode, the instantaneous downlink data rate on downlink RB $n_i$ is given by:

$$r_c^{n_i} = W_{RB} \log_2 \left( 1 + \frac{\sum_{s \in \mathcal{S}} P_{tx}^{n_i}\left(s_i, c\right) G\left(s_i, c\right) \delta\left(c, s_i\right)}{n_0 + I_c^{n_i}} \right) \tag{1}$$

where $P_{tx}^{n_i}\left(s_i, c\right)$ is the power transmitted by base station $s_i$ (macro or pico) on RB $n_i$ to user $c$, $G\left(s_i, c\right)$ is the channel gain between base station $s_i$ and user $c$, $n_0$ is the noise power, $W_{RB}$ is the RB bandwidth and $\delta\left(c, s_i\right)$ is a binary variable indicating whether $c$ is associated to $s_i$.

For VUEs, when this variable is set to 1, it implicitly means that they are using the network mode, and are associated to base station $s_i$. The second term in the denominator, $I_c^{n_i}$, is the interference experienced from other base stations in the network:

$$I_c^{n_i} = \sum_{s_j \in \mathcal{S} \wedge \mathcal{s_j \neq s_i}} P_{tx}^{n_i}\left(s_j, c'\right) G(s_j, c) \tag{2}$$

where $P_{tx}(s_j, c')$ denotes the power transmitted by base staion $j$, other than $i$, towards its allocated user $c'$ on RB $n_i$ and $G(s_j, c)$ is the channel gain between $s_j$ and user $c$.

Since D2D communications use uplink resources, we expect no interference on downlink RBs. In the uplink resources, the picture is a bit more complicated. On the one hand, we have the uplink cellular communications ongoing between UEs and their associated base stations, and on the other hand we have the concurrent D2D links between VUEs. Therefore, the uplink data rate for the arbitrary UE, $c$, on uplink RB, $n_j$, will be:

$$r_c^{n_j} = W_{RB} \log_2 \left( 1 + \frac{\sum_{s \in \mathcal{S}} P_{tx}^{n_j}\left(c, s_i\right) G\left(c, s_i\right) \delta\left(c, s_i\right)}{n_0 + I_c^{n_j}} \right) \tag{3}$$

In Eq. (3), $P_{tx}^{n_j}\left(c, s_i\right)$ indicates the transmitted power by $c$ towards base station $s_i$. The $I_c^{n_j}$ term here stands for the interference experienced at the macro base station from ongoing D2D communications:

$$\begin{aligned} I_c^{n_j} = \sum_{(v_i, v_j) \in \mathcal{V}} \Big[ & P_{tx}^{n_j}\left(v_i, v_j\right) G(v_i, s_i) \\ & + P_{tx}^{n_j}\left(v_j, v_i\right) G(v_j, s_i) \Big] \delta\left(v_i, v_j\right) \delta\left(v_j, v_i\right) \end{aligned} \tag{4}$$

We note here the double appearance of the $\delta$ variables and $P_{tx}$ parameters. As explained in detail below, a D2D link between two vehicles will only be established if the vehicles are mutually associated to each other. However, only one of the vehicles will be transmitting at a time, and that is indicated by correctly setting the $P_{tx}$ values. The cellular network resource allocation is orthogonal, therefore no interference will be caused between cellular users.

Finally, assuming a successful establishment of a D2D link, the instantaneous downlink data rate for a pair of VUEs, $(v_i, v_j)$, assuming $v_j$ is the receiving end, will be:

$$\begin{aligned} r_{v_j}^{n_j} = {} & W_{RB} \\ & \log_2 \left( 1 + \frac{P_{tx}^{n_j}\left(v_i, v_j\right) G\left(v_i, v_j\right) \delta(v_i, v_j) \delta(v_j, v_i)}{n_0 + I_{v_j}^{n_j}} \right) \end{aligned} \tag{5}$$

The interference at the receiving vehicle, $v_j$, will be two-fold, coming both from UEs transmitting in the uplink mode to the network and other ongoing D2D links:

$$\begin{aligned} I_{v_j}^{n_j} = & \sum_{c \in \mathcal{C}, \mathcal{s} \in \mathcal{S}} P_{tx}^{n_j}(c, s) G(c, v_j) \delta(c, s) \\ & + \sum_{(v_k, v_l) \in V \setminus (v_i, v_j)} P_{tx}^{n_j}(v_k, v_l) G(v_k, v_j) \delta(v_k, v_l) \delta(v_l, v_k) \end{aligned} \tag{6}$$

## IV. RL-MSAVP: RL-based Mode Selection and Vehicle Pairing

In D2D-enabled cellular networks, the devices may establish direct links between each other. In this context, vehicles in the network, which are also cellular users, can choose between

ISBN: 978-3-903176-71-3

the option to connect to the network (V2I) or directly to other vehicles (V2V), depending on the environment, i.e., channel conditions, as well as context.

This procedure, namely *mode selection* is applied for each vehicle to determine whether it will use the D2D mode with a nearby vehicle or connect to the network to exchange data traffic. As we saw in the previous section, the selected mode can have a significant impact both on the performance of the vehicular network as well as the cellular uplink communications.

In this work we focus in underlay V2X networks, in which vehicles opportunistically access the channel, without network coordination. In order to achieve a fully decentralized and network-independent solution to mode selection in underlay V2X networks, we consider that in addition to the mode selection, the vehicles must also be able to dynamically and intelligently choose their respective communicating peer. We assume that VUEs use Mode 2 resource allocation procedure based on their own sensing observations, to identify and use the cellular resources.

In this work, we propose the use of RL, specifically the use of Q-learning algorithm, to dynamically achieve this goal. RL is a popular machine learning technique that relies on feedback from the environment to improve its decision-making process. It presents low complexity, becoming very suitable to be used in practice.

To model our problem as an RL problem we must define four important components: the *agents*, the *state* and *action space*, and the *reward* which are used as feedback to update the decision-making metrics. The historical reward which is stored as Q-values, are used to determine the policy of each agent, i.e., the sequence of actions that each agent will undertake in response to its local observation of the environment.

The *agents* in our scenario reside at the individual vehicles that make independent decisions on whether to establish a D2D link or connect to the network, based solely on their own observations.

The *state space* is determined by the observations of the vehicle regarding the wireless environment. In our scenario, we consider two particular types of observations: the interference level observed in the downlink resources, and the interference level observed in the uplink resources. Therefore at each time step $t$, the state observed at vehicle $v$ is:

$$s_v^t = \begin{bmatrix} \sum_{n_i \in N_{DR}} I_v^{n_i} & \sum_{n_j \in N_{UR}} I_v^{n_j} \end{bmatrix} \tag{7}$$

Clearly, the state space in this scenario is continuous, and therefore infinite. To address the dimensionality issue of the state space, we apply a simple approach and quantize the observed values to make the state space discrete and finite. More complex approaches have used deep neural networks to approximate the value of the Q-function for a state-action tuple [13], which however introduces significant complexity to the system.

The *action space* is the set of actions that are available to the vehicle. We consider the mode selection and vehicle pairing to be an extension of the user association procedure. The action set of the vehicles is the set of user association decisions that are available to the vehicle. This includes the list of all potential serving points which can be either macro, pico base stations for V2I communication or other vehicles for V2V communications. Each vehicle has a different set of actions, we can define as:

$$\mathbf{A}_v = \begin{bmatrix} b_1 & 1 \\ \vdots & \vdots \\ b_k & 1 \\ v_1 & 2 \\ \vdots & \vdots \\ v_l & 2 \end{bmatrix} \tag{8}$$

where $b_1, \ldots, b_k$ denote the infrastructure nodes which are potential serving nodes, satisfying a received power threshold, while $v_1, \ldots, v_l$ are the vehicles that satisfy particular selection criteria and are potential communication peers for $v$. In this work we assume a simple distance threshold criteria for V2V eligibility. Therefore all vehicles within a certain distance from vehicle $v$ are admitted in the *action set* of $v$. However, this criteria may be further specified to include only vehicles that move in a certain direction, or vehicles with matching advertised resources when using Mode 2 sensing and resource-selection/reservation procedure [19].

The second column indicates the communication mode, 1 for V2I and 2 for V2V. Once an action $\hat{a}$ is selected by vehicle $v$, this is translated into the problem formulation correctly setting the association variable $\delta(v, \hat{a}[1])$ to 1 for the chosen endpoint, and to 0 for all other endpoints. At each time step, a user can only be associated to a single node, either in network or D2D mode. Therefore, the association variables must satisfy $\sum_{s \in \mathcal{S}} \delta(c, s) \leq 1, \forall c \in \mathcal{C}$ and $\sum_{s \in \mathcal{S}} \delta(v, s) + \sum_{v_i \in \mathcal{V}} \delta(v, v_i) \leq 1, \forall v \in \mathcal{V}$.

It is clear however that a problematic situation arises when a vehicle selects to associate to a vehicle that has chosen to associate to a third vehicle or the network. In such cases, we consider that the D2D link could not be established, therefore the vehicle will get no service. This will ultimately be reflected in the reward obtained for selecting a certain action $\hat{a}$. If, however, the selection of the vehicles is mutual, i.e, $\delta(v_i, v_j) = \delta(v_j, v_i) = 1$, then we will consider that a D2D link has been successfully established. Once a link has been successfully established, we consider that the use of the link between the two vehicles is tackled during the resource allocation process.

At each time step, only one of the endpoints may be transmitting while the other is receiving and this is reflected by setting the $P_t$ variables accordingly during the resource allocation procedure.

The *reward* is defined in terms of spectral efficiency (SE), which we have defined as the amount of data bits transmitted

ISBN: 978-3-903176-71-3

per RBs used:

$$\rho_v^t(s_v^t, a_v^t, s_v^{t+1}) = \frac{\sum_{n\in N_{DR}} r_v^n}{\eta_v^t} + \frac{\sum_{n\in N_{UR}}[r_v^n + \sum_{v_i\in V} r_{v_i}^n \delta(v,v_i)\delta(v_i,v)]}{\eta_v^t + \sum_{v_i\in V} \eta_{v_i}^t \delta(v,v_i)\delta(v_i,v)} \tag{9}$$

where $\eta_v^t$ stands for the number of RB-s allocated to vehicle $v$ at time $t$. Note that this value is determined by the underlying resource allocation procedure. For network communications we assume that all base stations apply a proportional fair algorithm, while for D2D communications we assume that the vehicles apply Mode 2 sensing and resource-selection procedure.

The first term in Eq. (9) stands for the downlink SE over the downlink resources and may be greater than 0 only if the vehicle chooses the network mode, and the second term stands for the SE over the uplink resources, either during uplink transmission or D2D communication. In case of D2D mode, a term has been added to the second part, to ensure that the entire amount of data transferred over the D2D link is counted towards the reward value, regardless of whether the vehicle is receiving or transmitting. For each $(s,a)$ tuple, the vehicle maintains a reward history, or Q-value, that is updated after action $a$ is selected while on state $s$, as follows:

$$Q_v(s,a) = (1-\alpha^t)Q_v(s,a) + \alpha^t[\rho_v^t(s^t, a_v^t, s_v^{t+1}) - \bar{\rho}_v^t + \nu \max_{a'\in\mathbf{A}_v} Q(s_v^{t+1}, a')] \tag{10}$$

where $\bar{\rho}_v^t$ is the average reward for vehicle $v$ at time $t$. Parameters $\alpha$,$\beta$ and $\nu$ are algorithm-specific parameters which can be tuned [22]. The algorithm is executed in a decentralized manner by each vehicle, in two phases, referred to as the *learning phase* and *frozen phase*. During the learning phase, at each time step $t$, vehicle $v$ while at state $s_v^t$, chooses either the action $a$ that maximizes the $Q_v(s_v^t, a)$ value, with a certain probability, or picks randomly one of the other possible actions. This is done for exploratory purposes, however the probability of exploration diminishes as time progresses. After the selected action $a$ is enacted and the system progresses to state $s_v^{t+1}$ the immediate reward is calculated and the Q-value updated according to Eq. (10).

During the frozen phase, the vehicles simply choose the action that maximizes the $Q_v$ value, and the $Q_v$ values are not updated. Since the system is dynamic and the environment changes due to vehicle mobility and channel conditions, the two phases are conducted repeatedly during the simulation period. Since the algorithm is fully decentralized, in general, the vehicles do not need to synchronize their decision making processes. Therefore the time step at which the decisions are made, as well as the length of the learning and frozen phases may be individually configured by the vehicles.

## V. Simulation Results

We evaluate the performance of our scheme through numerical simulations on the MATLAB platform. The simulation

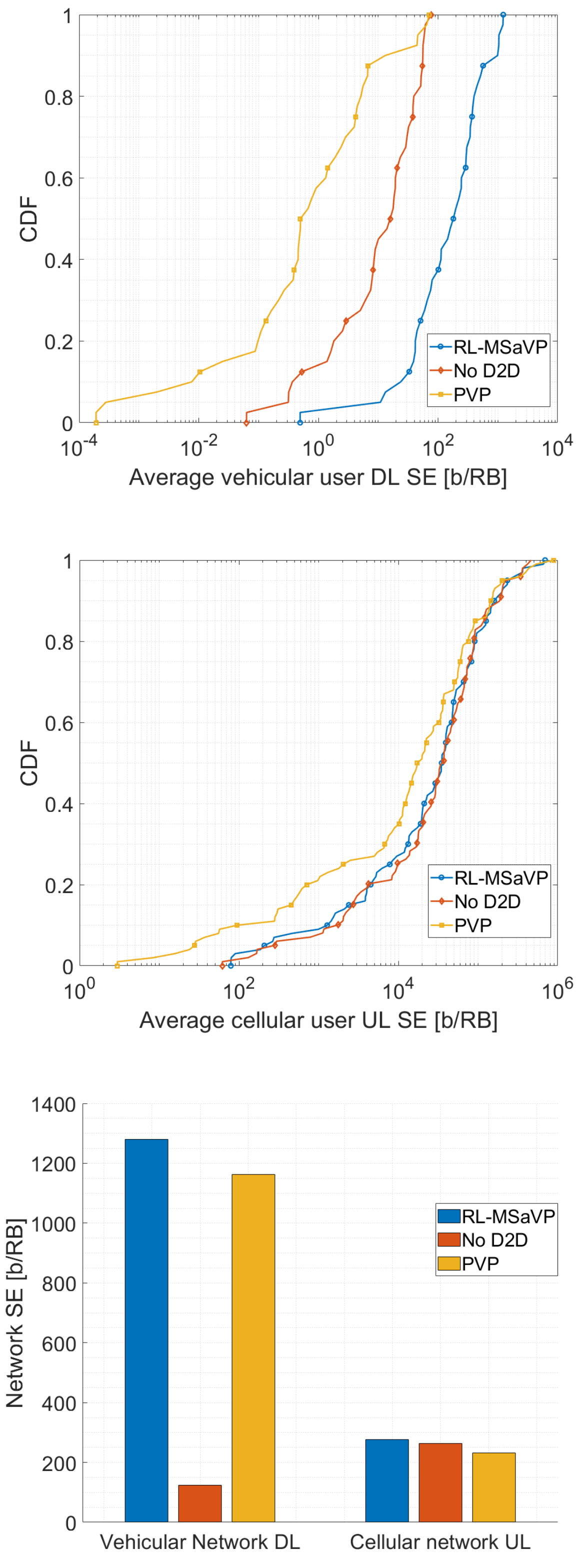

Fig. 2: Spectral efficiency (SE): cumulative distribution function (CDF) of the average SE of individual vehicles (top) and cellular users in the uplink (middle). The network SE of the vehicular and the cellular network (bottom).

ISBN: 978-3-903176-71-3

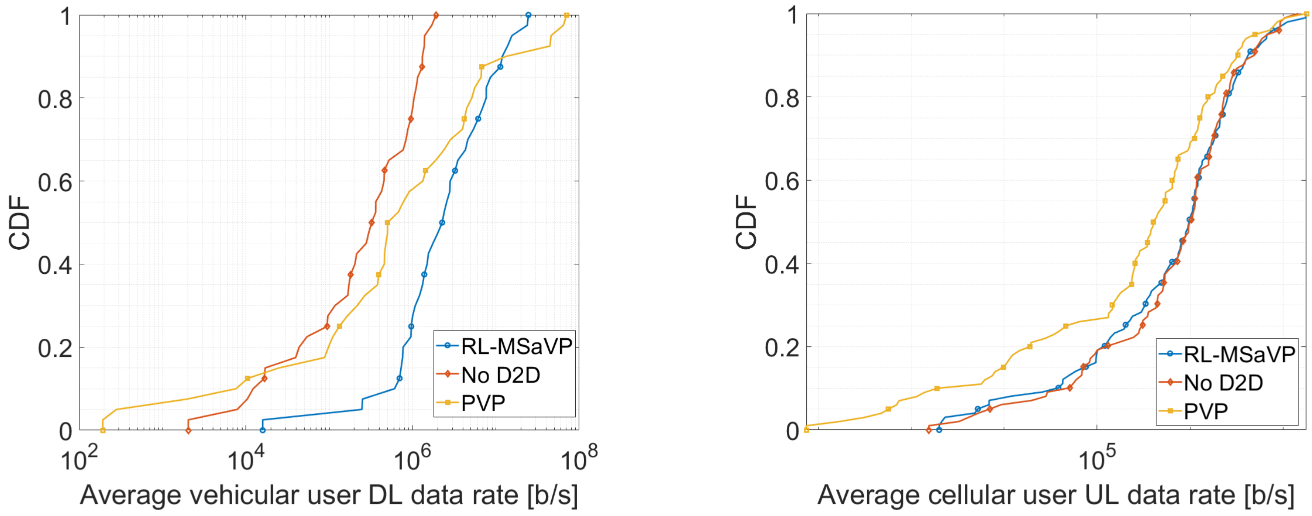

Fig. 3: Cumulative distribution function (CDF) of the average VUE data rate in the downlink direction (left) and average cellular UE data rate in the uplink direction (right).

TABLE I: Simulation Parameters

| Parameter | Value |
|---|---|
| Number of cellular users, C | 100 |
| Number of vehicle users, C | 40 |
| Number of RBs | 50 |
| Subcarrier spacing | 15 kHz |
| Maximum user/vehicle transmit power $P_t^{max}$ | 24 dBm |
| Maximum macro/pico transmit power | 43/30 dBm |
| Road segment length | 800m |
| Road segment width | 25m |
| V2V distance threshold | 50m |
| Simulation period | 4s (4000 timeslots) |
| Learning phase/frozen phase duration | 0.5s (500 timeslots) |
| Noise level | $-174$ dBm/Hz |

parameters, based on ITU recommnendations [23] are shown in Table 1.

We simulate a network consisting of a single macro base station, 9 pico base stations and 100 cellular users distributed uniformly within the coverage area. In two intersecting road segments, we place 40 vehicles randomly. Their movement is determined by their position on the road. We assume that the network BSs allocate their uplink and downlink RBs using the proportional fair algorithm. The UEs and VUEs choose the network BS according to the standard user association procedure, which relies on the strongest received pilot signal. For VUEs, the network BS is added to the list of potential association nodes, which also may include other VUEs. To mitigate inter-tier interference we also apply cell range expansion (CRE) with 10 dB offset. The time slot is 1ms, which corresponds to 1 subframe in the 5G numerology with 15 kHz subcarrier spacing.

When using the RL-based approach, the vehicles decide which mode and node they will use for communication at every time slot, and in the case when V2V mode is selected which vehicle they will choose for establishing V2V link. Once each user and vehicle determines its user association scheme, we run the simulation and obtain the actual instantaneous data rates achieved by the cellular and vehicular users as well as the number of RBs used. These are used to calculate the immediate rewards and update the Q-values. The state values are discretized in logarithmic scale using 1 dB quantization steps. The small-scale fading variations of the channel are randomly generated every time slot.

We compare our RL-based mode selection and vehicle pairing scheme (denoted as 'RL-MSaVP') with two other scenarios:

1) The network does not support D2D communications, therefore all vehicles are treated like regular cellular users (denoted as 'No D2D'). All vehicles associate to the infrastructure node that maximizes the received power level.
2) All vehicles use strictly D2D communications and the communicating vehicle pairs are predetermined by the network according to their best mutual received power (denoted as 'PVP - Predetermined Vehicle Pairing').

First, we look at the spectral efficiency (SE) of the individual vehicles, which is the performance indicator each vehicle aims to improve individually. The cumulative distribution function of the average vehicle SE in the downlink direction is shown in Fig. 2 (top). Note that we consider that D2D links are used to offload non-safety downlink traffic, therefore all data exchanged during D2D communications are categorized as downlink, although they occur in the uplink resources. As expected, RL-MSaVP outperforms the two approaches by a significant margin. In the PVP scheme, in which vehicles are using D2D mode only, they are a priori paired to their

ISBN: 978-3-903176-71-3

best match in terms of received power, however this does not guarantee the match is most efficient in terms of SE. We also want to look at how the ongoing D2D connections affect the uplink cellular communications. In Fig. 2 (middle) we have shown the individual average SE of the UEs in the uplink direction. We note that RL-MSaVP affects the SE of the uplink communications only marginally, while indiscriminate use of D2D links, has a notable degrading effect. This is reflected also in the performance of SE at the network level which is shown in Fig. 2 (bottom). Although the approach is decentralized, the proposed approach still manages to outperform at the network level as well, with RL-MSaVP guaranteeing a higher network SE both for the vehicular network in the downlink, and the cellular network in the uplink.

Next, we also look at the performance of the scheme for another QoS indicator, which is the average data rate obtained by the vehicular users in the downlink and the cellular users in the uplink (Fig. 3). Although the RL-MSaVP algorithm does not optimize this indicator explicitly, we note that RL-MSaVP still manages to guarantee significantly higher average data rates than both schemes, at least for 90% of the vehicular users, as shown in Fig. 3 (left). In addition, it is able to achieve that, while causing minimal degradation to cellular uplink communications, as we see that the CDF curve of the average uplink data rate almost matches the 'No D2D' scheme (Fig. 3 (right)). The harmful effect of the V2V interference on the uplink communications, even when vehicles are matched by the network, can be seen in Fig. 3 (right) where we notice a significant drop of average data rate for at least 90% of cellular uplink users when applying the 'PVP' scheme.

Finally we look at algorithm behaviour in a real-time simulation of the environment. Fig. 4 (left) presents the time evolution of the average vehicular network SE during the simulation run. We recall here that the algorithm is executed in two phases repeatedly: the learning and frozen phases which are applied in real-time. Hence we notice the sudden drops in SE and then the ascend of the network SE as the algorithm updates its policies. RL-MSaVP is better equipped to respond to vehicle mobility compared to a fixed solution such as PVP, as we can see in Fig. 4 (left) that SE under PVP drops during the simulation execution time as vehicles move. We note that 'No D2D' scheme is less affected by the vehicular mobility, which is expected. Fig. 4 (right) also presents the evolution of the individual average rewards of the vehicles. We note here, that although the individual average rewards, $\bar{\rho}$, experience lows and highs as the algorithm re-adapts to a changing environment, the averaged average rewards over the set of vehicles maintains a stable level, which supports the earlier findings that the behavior of RL-MSaVP at the network level is better than expected, despite the decentralized implementation.

## VI. Conclusion and future work

In this work we have presented RL-MSaVP, a RL based scheme for spectral efficient mode selection and vehicle pairing for a V2X network underlaying a multi-tier cellular network. The goal was to dynamically and independently solve the problem of selecting a communication mode and communicating peer in a manner which ultimately optimized the spectral efficiency of the system.

The performance of the solution is evaluated in a two-tier cellular network operating in an urban environment, with realistic modeling of vehicular mobility, using numerical simulations, and compared to two other benchmark solutions. The performance of our approach is evaluated in terms of spectral efficiency and average user data rate. The results show that the proposed technique can improve the performance of the individual VUEs when selecting the communication mode, by ensuring higher SE for 90% of the VUEs, and furthermore it is able to do while causing minimal disruption to the cellular uplink communications. In addition, we also record an estimated 12 % improvement in terms of SE at the network level, despite the decentralized implementation.

Future work will focus on further tuning the algorithm design and parameters to optimize its performance, as well as the individual adjustment of the decision time step and learning phase duration. In addition, we plan to investigate the application of the approach in more complex networks which feature mobile base stations such as UAVs and mmWave communications.

## Acknowledgment

This work was supported by the Ministry of Education, Science and Innovation of Republic of Kosovo, through the annual scientific grants project.

## References

[1] M. H. C. Garcia et al., "A Tutorial on 5G NR V2X Communications," in IEEE Communications Surveys & Tutorials, vol. 23, no. 3, pp. 1972-2026, thirdquarter 2021, doi: 10.1109/COMST.2021.3057017.

[2] Y. Zhang, F. Tian, B. Song, and X. Du, "Social vehicle swarms: A novel perspective on socially aware vehicular communication architecture," IEEE Wireless Commun., vol. 23 no. 4, pp. 82–89, Aug. 2016.

[3] M. Harounabadi, D. M. Soleymani, S. Bhadauria, M. Leyh and E. Roth-Mandutz, "V2X in 3GPP Standardization: NR Sidelink in Release-16 and Beyond," in IEEE Communications Standards Magazine, vol. 5, no. 1, pp. 12-21, March 2021.

[4] F. Jameel, Z. Hamid, F. Jabeen, S. Zeadally, and M. A. Javed, A survey of device-to-device communications: Research issues and challenges, IEEE Commun. Surveys Tuts., vol. 20, no. 3, pp. 21332168, 3rd Quart., 2018.

[5] A. Hegde, A. Festag, Mode Switching Strategies in Cellular-V2X, IFAC-PapersOnLine, Volume 52, Issue 8, 2019, Pages 81-86, ISSN 2405-8963, https://doi.org/10.1016/j.ifacol.2019.08.052.

[6] X. Zhang, M. Peng, S. Yan and Y. Sun, "Deep-Reinforcement-Learning-Based Mode Selection and Resource Allocation for Cellular V2X Communications," in IEEE Internet of Things Journal, vol. 7, no. 7, pp. 6380-6391, July 2020, doi: 10.1109/JIOT.2019.2962715.

[7] H. D. R. Albonda and J. Pérez-Romero, "A New Mode Selection and Resource Reuse Strategy for V2X in Future Cellular Networks," 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring), Antwerp, Belgium, 2020, pp. 1-6, doi: 10.1109/VTC2020-Spring48590.2020.9129454.

[8] Luca Lusvarghi, Alejandro Molina-Galan, Baldomero Coll-Perales, Javier Gozalvez, Maria Luisa Merani, A comparative analysis of the semi-persistent and dynamic scheduling schemes in NR-V2X mode 2, Vehicular Communications, Volume 42, 2023, 100628, ISSN 2214-2096,https://doi.org/10.1016/j.vehcom.2023.100628.

ISBN: 978-3-903176-71-3

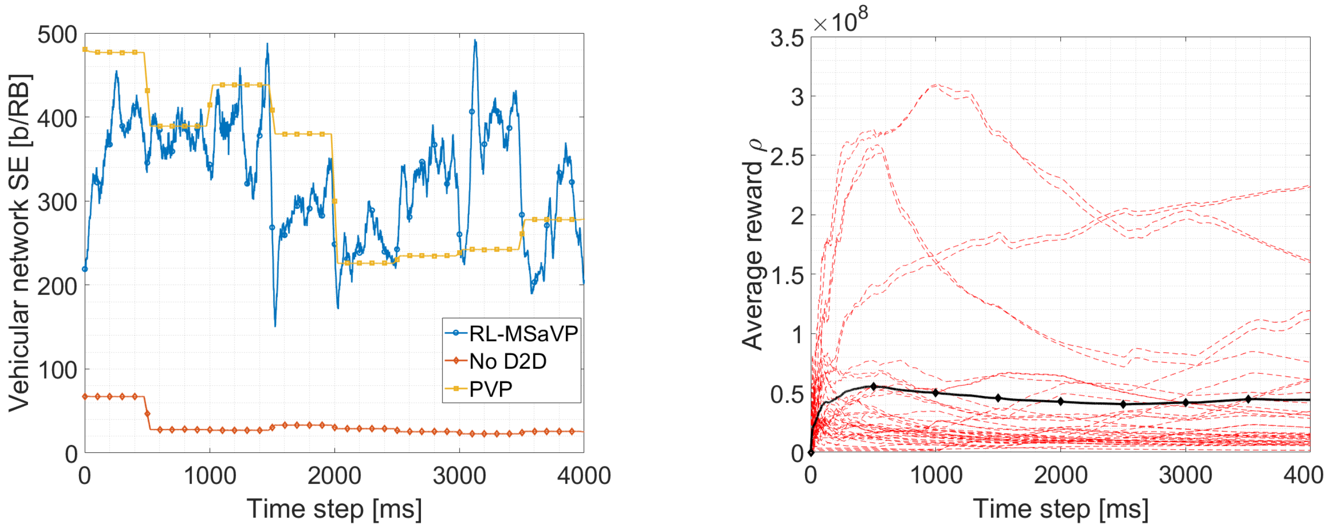

Fig. 4: Time evolution of the average SE of the vehicular network during the simulation period (left) and time evolution of the vehicle average rewards $\bar{\rho}$ (right). The individual average rewards are shown in dashed red lines, while the thick black line represents the averaged $\bar{\rho}$.

[9] E. E. Gonzalez, Y. Estrada, D. Garcia-Roger and J. F. Monserrat, ”Performance Evaluation and Optimal Management of Mode 2 V2X Communications in 5G Networks,” in IEEE Access, vol. 11, pp. 128810-128825, 2023, doi: 10.1109/ACCESS.2023.3333680.

[10] A. Molina-Galan, L. Lusvarghi, B. Coll-Perales, J. Gozalvez and M. L. Merani, ”On the Impact of Re-Evaluation in 5G NR V2X Mode 2,” in IEEE Transactions on Vehicular Technology, vol. 73, no. 2, pp. 2669-2683, Feb. 2024, doi: 10.1109/TVT.2023.3318235.

[11] M. Sana, A. De Domenico, W. Yu, Y. Lostanlen and E. Calvanese Strinati, ”Multi-Agent Reinforcement Learning for Adaptive User Association in Dynamic mmWave Networks,” in IEEE Transactions on Wireless Communications, vol. 19, no. 10, pp. 6520-6534, Oct. 2020.

[12] J. Moon, S. Kim, H. Ju and B. Shim, ”Energy-Efficient User Association in mmWave/THz Ultra-Dense Network via Multi-Agent Deep Reinforcement Learning,” in IEEE Transactions on Green Communications and Networking, vol. 7, no. 2, pp. 692-706, June 2023.

[13] D. Zhao, H. Qin, B. Song, Y. Zhang, X. Du and M. Guizani, ”A Reinforcement Learning Method for Joint Mode Selection and Power Adaptation in the V2V Communication Network in 5G,” in IEEE Transactions on Cognitive Communications and Networking, vol. 6, no. 2, pp. 452-463, June 2020.

[14] C. Kai, X. Meng, L. Mei and W. Huang, ”Deep Reinforcement Learning Based User Association and Resource Allocation for D2D-enabled Wireless Networks,” 2021 IEEE/CIC International Conference on Communications in China (ICCC), Xiamen, China, 2021, pp. 1172-1177.

[15] Q. Guo, F. Tang and N. Kato, ”Federated Reinforcement Learning-Based Resource Allocation in D2D-Enabled 6G,” in IEEE Network, vol. 37, no. 5, pp. 89-95, Sept. 2023.

[16] Dan Wang, Hao Qin, Bin Song, Ke Xu, Xiaojiang Du, Mohsen Guizani, “Joint resource allocation and power control for D2D communication with deep reinforcement learning in MCC”, in Physical Communication, vol. 45, 2021

[17] J. Tan, Y.-C. Liang, L. Zhang and G. Feng, “Deep reinforcement learning for joint channel selection and power control in D2D Networks,” in IEEE Trans. Wireless Commun., vol. 20, no. 2, pp. 1363–1378, 2021.

[18] 3GPP Technical Specification. ”5G; NR; User Equipment (UE) radio transmission and reception; Part 1: Range 1 Standalone (3GPP TS 38.101-1 version 17.5.0 Release 17)”, May 2022.

[19] Erik Dahlman, Stefan Parkvall, Johan Sköld, ”5G/5G-Advanced (Third Edition)”, Academic Press, 2024.

[20] 3GPP Technical Report. “5G;Study on channel model for frequencies from 0.5 to 100 GHz (3GPP TR 38.901 version 14.3.0 Release 14)”, 2018.

[21] F. Malandrino, Z. Limani, C. Casetti and C. -F. Chiasserini, ”Interference-Aware Downlink and Uplink Resource Allocation in HetNets With D2D Support,” in IEEE Transactions on Wireless Communications, vol. 14, no. 5, pp. 2729-2741, May 2015.

[22] A. Gosavi. Simulation-Based Optimization: Parametric Optimization Techniques and Reinforcement Learning, Springer, New York, NY, Second edition, 2014.

[23] ITU-R. Guidelines for evaluation of radio interface technologies for Guidelines for evaluation of radio interface technologies for IMT-2020; 2017. Report ITU-R M.2412-0.

ISBN: 978-3-903176-71-3