

Joint Routing and Wireless Resource Allocation in Multihop LTE-D2D Communications

Safwan Alwan*, Ilhem Fajjari[†] and Nadjib Aitsaadi[‡]

*University Paris-Est, LiSSi EA 3956, UPEC, F94400, Vitry-sur-Seine, France

[†]Orange Labs, F92320, Chatillon, France

[‡]University Paris-Est, LIGM-CNRS UMR 8049, ESIEE Paris, F93160, Noisy-le-Grand, France

Emails: safwan.alwan@univ-paris-est.fr, ilhem.fajjari@orange.com, nadjib.aitsaadi@esiee.fr

Abstract—5G aims to maximize the data rate and to handle the billions of video, voice, data and IoT flows. For this reason, the macro-cells will be very congested and may fail to satisfy the end-users. In this context, data offloading scheme is conceived to route intra-cell traffic among the D2D-enabled user equipments reusing wireless uplink resources and thus increasing the overall spectral efficiency. In this paper, we address the joint routing and OFDMA resource allocation problem in D2D network. To do so, first we formulate the problem as Mixed Integer Linear Programming. The model takes into account factors that limit spectrum reuse as well as other LTE-D2D technology constraints such as: half-duplex operation and contiguity in resource block allocations. Then, we propose a novel scheme named Joint Routing and Wireless allocation in D2D communications (JRW-D2D) which is based on the branch-and-cut algorithm. In order to gauge the effectiveness of our proposal, we implement the standard LTE-D2D protocol stack, including our scheme JRW-D2D, in the NS-3 network simulator. The results obtained are very promising in terms of reliability, ratio of admitted D2D flows and latency in comparison to other basic one-sided optimal strategies including an interference-aware heuristic scheme.

Index Terms—Multihop routing, Resource allocation, Offloading, Device-To-Device LTE-D2D, Optimization.

I. INTRODUCTION

There is no denying that the Fourth Generation (4G) of mobile cellular network, Long Term Evolution (LTE), held the promise of higher data rate and enhanced the Quality of Service (QoS). But, the growth of video-centric and social media services has led to the explosion of traffic demand. In addition, Internet of things will exponentially increase the number of flows in the cellular network. Consequently, current operator infrastructures struggle to accommodate the required network resources and link capacities. This trend is set to continue and recent statistics highlight that the number of connected devices is estimated to reach 50 billion by 2020 while the mobile data traffic is expected to grow to reach 49 exabytes per month by 2021 [1].

Therefore, discussions of a new standard have taken place in both industry and academia to design the Fifth Generation (5G) mobile cellular network architecture. The main objective of 5G is to ensure the QoS satisfaction of the different applications and to deal with diverse deployments in terms of available resources and connected devices requirements. In this context, 5G puts forward disruptive technologies making use of i) massive MIMO and millimeter-Wave antenna systems, ii) Multiple Radio Access Technologies (Multi-RAT), iii) small cells deployment and iv) advanced Device-to-Device (D2D) communications. All these techniques aim to increase the capacity of networks in order to handle large number of connections and data volume at high throughput and very low latency.

The main idea behind D2D is to enable direct communications between devices in close proximity and thus bypassing macro base-stations (eNodeB). D2D was incorporated in LTE-A to increase the spectral efficiency of cellular systems and to support new use cases such as: i) public safety scenar-

ios, ii) device-discovery for commercial applications, iii) D2D-network relays, etc. D2D is also one pillar of 5G architecture enabling operators to ensure an extended and controlled connectivity while reducing the networks cost thanks to the traffic offloading solutions. In doing so, the data plane is moved from operator's infrastructure (i.e., E-UTRAN and EPC) to end-users' devices (i.e., UE). However, the control plane is managed by the operator and hosted in E-UTRAN. This will alleviate the infrastructure's load while enabling large numbers of simultaneous connections with better QoS.

D2D raises several design challenges [2], [3] such as: coexistence with conventional communications mode (macro-cell), spectrum reuse and resource allocation, mode-switching, extending single-hop scenarios to multihop ones, etc. In this paper, we address the routing and wireless resource allocation problems in D2D communications. Multihop D2D seeks to enhance the utility of D2D systems by increasing communication range and reducing the load in the operator's infrastructure. Multihop D2D system must adopt various policies with respect to the routing, resource allocation, interferences: intra-mode (i.e., D2D links) and inter-mode (i.e., D2D and conventional communications). Note that a sidelink communication (i.e., D2D) uses the same physical resource (transceiver and spectrum) of uplink communication. That means, UE cannot simultaneously run both sidelink and uplink communications. In addition, UE cannot simultaneously transmit and receive in the sidelink. Consequently, each link in the D2D path is half-duplex and only non-critical (in terms of latency and bandwidth) traffic can be handled. We formulate the joint routing and resource allocation problem of D2D communications while considering: i) contiguity of OFDMA resource block allocation, ii) interference and iii) half-duplex mode of operation in LTE-D2D as a Mixed Integer Linear Programming (MILP) problem. The objective is to maximize the bandwidth of each flow (i.e., best effort).

To solve the above problem, we propose novel scheme, based on the branch-and-cut algorithm [4], named **Joint Routing and Wireless allocation in D2D communications** (JRW-D2D). In this paper, we assume a dense deployment of UEs in delimited area such as stadium. Consequently, the UEs are not mobile. It is worth noting that the routes set up for flows are *semi-static* paths. In other words, each path is maintained for the whole period of communication to avoid excessive signaling to reconfigure D2D links. On the other hand, resource allocations are dynamically executed every assignment interval to cater for flow arrivals and departures. To assess the performance of our proposal JRW-D2D, we implemented the LTE-D2D protocol stack in the NS-3 network simulator to support this standard. In doing so, the whole protocol stack is simulated and hence the conclusions will be more significant than the numerical simulations. The results obtained demonstrate the effectiveness of JRW-D2D in terms

of the optimality, ratio of admitted D2D flows and latency. In addition, we compared our proposal to other basic one-sided optimal strategies and an interference-aware heuristic scheme. One-sided optimal strategy is one which is optimal only in one sense either in terms of routing or in terms of resource block allocation.

The remainder of this paper is organized as follows. In section II, we will present an overview of existing literature that deals with the issue of multihop D2D communications. In section IV, we will describe in details our system model for the offloading application and how we formulate the decision problem as a MILP model that includes routing and resource block allocation. Then, in section V, we will describe our proposal JRW-D2D used to solve the underlying problem. Next, in section VI, we will present our evaluation methodology and network simulation results. Finally, Section VII will conclude the paper.

II. RELATED WORK

D2D wireless communications in cellular networks is an extremely challenging paradigm that has aroused the interest of both industry and academia. In this section, we summarize the most relevant related work that helped us to have an insight into the multihop D2D routing and resource allocation problems.

In [5], the authors propose D2D multihop communications in cellular networks for a public safety scenario under partial coverage. Using a homegrown 3GPP standard compliant system level, the authors claim to have demonstrate improvements in both energy and spectral efficiency compared to conventional communications. However, the routing algorithm was not introduced in their public safety scenario. In fact, only predefined routes are used in to allow one far-away UE to reach an active base-station using the other UEs as relays.

In [6], the authors put forward a two-stage method to find multihop D2D paths under a limit on the maximum interference incurring at conventional mobile users. Based on numerical simulations, significant improvements in throughput can be achieved using multihop paths compared to single-hop D2D communication. However, the proposed method is highly generic. Indeed, only one single assumption is considered by the authors to apply their approach: downlink resources are shared by D2D and conventional communication.

In [7], the authors study the optimal transmission scheduling and congestion control in multihop D2D communications underlaying cellular networks while taking into account: i) interference situation for D2D mode and conventional cellular mode and ii) QoS requirements of each traffic flow. Making use of Lyapunov optimization theory, the formulation considers the following problems in order to solve the global problem in a sub-optimal way: i) end-to-end rate control, ii) joint routing and channel assignment, and iii) power allocation. It is worth pointing out that the proposed algorithm takes into account also the stability of queues in forwarding devices since it adopts a dynamic routing where decision is done on per-time-slot basis. Remarkably, the proposed algorithm also makes a number of assumptions relevant to LTE-D2D: i) D2D links share the uplink spectrum and ii) half-duplex nature of D2D links. However, the spectrum allocation in terms of resource block is not considered which renders it less realistic in the view of LTE-D2D standard.

In our work, we address the joint optimization of resource block allocation and routing for multihop communications. Unlike [7], we adopt a semi-static routing where path establishment takes into account the current state of interfering

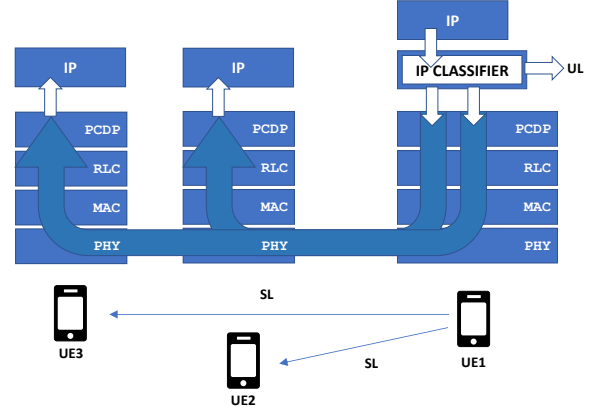


Fig. 1. LTE-D2D protocol stack for direct communications

links, but the path is held for the whole period of communication. We also model the allocation problem to the resource block level taking into account the fact that they are allocated in contiguous manner (3GPP uplink constraint). Besides, we notice that existing literature on multihop D2D communications shows varying degree of relevancy to LTE-D2D standard and lack of proposal validation using network simulators due the support for D2D standards. To cope with this limitation, we implemented in NS-3 the full 3GPP LTE-D2D protocol stack to evaluate the performance of our proposal.

III. LTE-D2D ARCHITECTURE

To support LTE-D2D, an enhanced user equipment (UE) contains an additional protocol stack besides the conventional one. This new LTE-D2D stack provides the so-called proximity-based services (ProSe) to the upper layer(s) [8]. ProSe includes: i) Direct Discovery: a service whereby a UE is able to detect and identify other UEs in its proximity, ii) Direct Communications: UEs can directly communicate with each other bypassing the cellular infrastructure, and iii) UE-to-Network Relay: remote UE uses another UE as a relay in the network.

From an upper level perspective, ProSe are carried over a new type of wireless link besides the conventional ones (i.e., DownLink (DL) and UpLink (UL)). This lateral link between UEs is called SideLink (SL). In LTE-D2D, SL is configured to use the same frequency resources as UL to increase the overall spectral efficiency. It also reuses much of UL structure and hardware to add another efficiency dimension. From the lower layers perspective, SL presents its direct communication services to the upper layers in terms of no-feedback SL Radio Bearers (SLRBs). This is done to present a uniform support for both unicast and multicast IP communications [8].

A. LTE-D2D protocol stack for direct communications
Fig. 1 depicts the LTE-D2D stack to support direct communications. Hereafter, we provide a brief top-down description.

1) PDCP/RLC

Similar to their counterparts in the conventional LTE communication stack, Packet Data Convergence Protocol and Radio Link Control layers provide IP packet segmentation, header compression and security procedures. A single SLRB is identified by pair of PDCP/RLC entities connected in tandem at the source UE and the corresponding pair(s) at the destination UE(s). At the interface with incoming packets from the upper IP layer resides an IP-flow classifier that directs each IP packet to its corresponding SLRB PDCP/RLC entities.

2) MAC

It serves the upper layers by transmitting Transport Block (TB) composed of RLC Protocol Data Units (PDU) from

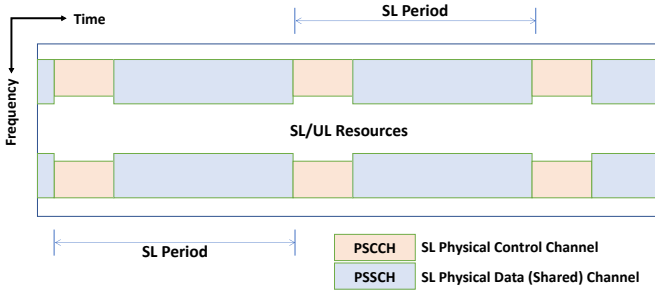


Fig. 2. SL channel structure in LTE-D2D (direct communications only)

possibly several SLRB bearers as long as they have the same destination. Each TB is identified by its layer L2 identifiers namely i) the source *ProSe-UE-ID* and ii) the destination *ProSe-L2-Destination-ID*. A TB is transmitted when a new SL transmission opportunity arrives while Hybrid ARQ (HARQ) operations in LTE-D2D is restricted to blind retransmissions (i.e., with no feedback) to increase the reliability. Hence, each TB is further retransmitted three times in the subsequent transmission opportunities with different redundancy versions.

3) PHY

Similar to UL, SL transmission uses the SC-OFDM modulation format using the grid of resource blocks. The latter occupies a subframe (i.e., TTI) which lasts for 1 ms and a bandwidth of 12 sub-carriers (180 KHz) in frequency domain. However, unlike UL, SL allocations are organized in longer periodic intervals called SideLink control Periods (SL-Periods), which can be configured between 40 and 320 subframes in length. As depicted in Fig. 2, a SL-Period starts with a control part followed by a data part. However the information, on which subframes and RBs are actually available for the operation, is conveyed by a configuration parameter called a *resource pool*.

A UE interested in SL reception continuously scans the configured resource pool(s) (the control part of SL periods) to check for incoming data. On the other hand, a UE wanting to transmit on SL transmission part, may be configured to go ahead and autonomously selects a subset RBs to use for transmission from a resource pool configured for this mode of operation. Even with in-coverage scenarios, the eNodeB may configure this autonomous mode for its UEs. Another possible option is to configure *scheduled-resources pool* to be used in the scheduled mode which gives the eNodeB a finer control over resource allocation. In this mode, a grant from the eNodeB to the UE determines which RBs and subframes to be used by the UE to transmit on SL. In this paper, we assume the latter mode of resource allocation. In doing so, the eNodeB has a total control on the resource allocation.

B. Synchronous operation of SL

Transmission and reception on SL are synchronous operations that require a common synchronization reference for all parties in the system. With in-coverage scenarios, where a UE is inside the coverage zone of an eNodeB, the UE synchronizes its SL operation to the timing of the related macro-cell which acts as a synchronization reference. Further procedures and provisions are given in the standard allowing some UEs to relay timing reference to extend the synchronization zone even under out-of-coverage scenarios [8].

C. Half-duplex operation of SL

As per LTE-D2D standard, the duplex mode of SL is half-duplex meaning that a UE can not simultaneously both listen and transmit on SL. However, the rules for role-switching are not specified and left to the application under consideration. In

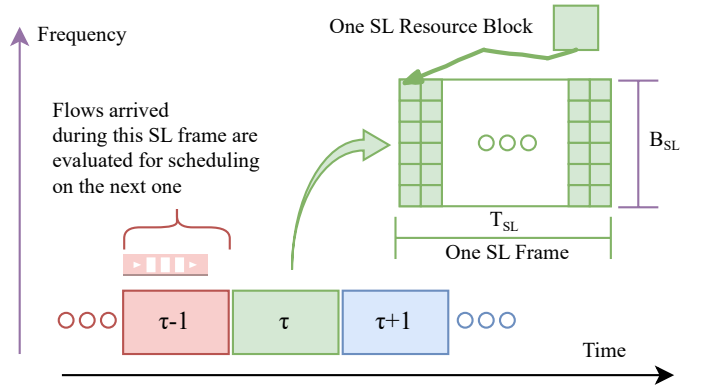


Fig. 3. Sidelink frame structure and scheduling

addition, UEs connected to the eNodeB (i.e., macro-cell) are required by the standard to give their UL transmissions higher priority over SL transmissions since both compete for the same SC-OFDMA transmitter. So, whenever there are UL data or reports (e.g., SRS, CQI, etc.), an ongoing SL subframe, if any, must be dropped. These properties make SL transmissions more opportunistic and intermittent and less reliable than UL.

IV. SYSTEM MODEL AND PROBLEM FORMULATION

Our system model considers N UEs inside the coverage zone of a single LTE-A eNodeB. These UEs, which are supposed to support the LTE-D2D protocol, are willing to offload the intra-cellular traffic between them when commanded to do so by the central controller in the eNodeB. We also assume that these UEs are quasi-stationary nodes. The eNodeB supervise the offloading operation over this D2D subnetwork by continuously allocating radio resources in every SL frame with decision instants given by:

$$t = \tau \times T_{SL} \quad \text{for } \tau = 0, 1, 2, 3, \dots$$

where T_{SL} is the duration of SL frame. The SL frame, or the SL control period in LTE-D2D terminology, is the scheduling time unit in SL which spans multiple one-millisecond time slots (i.e., multiple TTIs). Fig. 3 illustrates the structure of SL frame. The eNodeB models the D2D topology as a symmetric directed graph $\mathbb{G} = (\mathcal{V}, \mathcal{E})$. The set of vertices $\mathcal{V}$ and the set of edges $\mathcal{E}$ represent the UE nodes and the links between the UEs (i.e. SLs) respectively. Note that a link in topology is formed, and hence an edge exists in $\mathbb{G}$, only when the achieved SNR is greater than a threshold γ_{TOPO} . This means that $\mathbb{G}$ is not (fully-)connected in the general case and can be expressed as a union of connected subcomponents: $\mathbb{G} = \mathbb{G}^1 \cup \mathbb{G}^2 \cup \dots \cup \mathbb{G}^C$.

The problem of finding an offloading path for a flow $f^k \in \mathcal{F}$, whose source and destination are s^k, d^k respectively, can be formulated as follows: We introduce for each link e_{ij} a binary variable x_{ij} to indicate whether it is selected to be a part of some route. We also introduce for each node v_n a binary variable A_n^k that indicates whether it is associated with the flow f^k . In this formulation the offloading path for f^k is defined by the set of $\mathbb{P}^k \subseteq \mathcal{E}$:

$$\mathbb{P}^k = \{e_{ij} \in \mathcal{E} \mid x_{ij} = 1 \wedge A_i^k = A_j^k = 1\} \quad (1)$$

However, in order for equation (1), to meaningfully define a path, the solution space must respect some constrained defined in the following.

First we impose that nodes are exclusive for concurrent flows. In other words, a node can route at most one flow at a time. Formally, this constraint is introduced as:

$$\forall v_n \in \mathcal{V}, \sum_{f^k \in \mathcal{F}} A_n^k \leq 1 \quad (2)$$

In addition, if a node is associated with some flow it must have exactly one *incoming* links selected except at the source where there is none. This is formally imposed as:

$$\forall v_n \in \mathcal{V}, \sum_{e_{ij} \in \mathcal{E} | j=n} x_{ij} = \sum_{f^k \in \mathcal{F}, v_n \neq s^k} A_n^k \quad (3)$$

Similarly, if a node is associated with some flow it must have exactly one *outgoing* links selected except at the destination where there is none, or:

$$\forall v_n \in \mathcal{V}, \sum_{e_{ij} \in \mathcal{E} | i=n} x_{ij} = \sum_{f^k \in \mathcal{F} | v_n \neq d^k} A_n^k \quad (4)$$

Also, to ensure that node association is consistent with link selection, the following constraint imposes that the ends of a selected links are associated with the same flow:

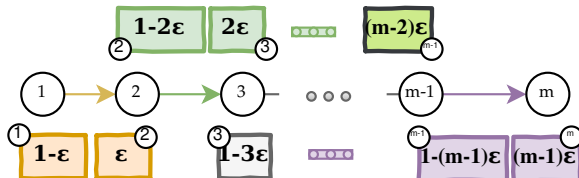
$$\forall f^k \in \mathcal{F}, \forall e_{ij} \in \mathcal{E}, x_{ij} - 1 \leq A_j^k - A_i^k \leq 1 - x_{ij} \quad (5)$$

It is straightforward to see that, if some flow f^k is decided to be admitted, which is indicated by $A_{s^k}^k = 1$, then $\mathbb{P}^k$, as defined in equation (1) and under Constraints (2) to (5), must contain only one simple path, which starts from s^k to d^k . However these constraints do not rule out superfluous links (and nodes) from appearing in $\mathbb{P}^k$ forming *simple isolated* cycles between them.

To exclude these isolated simple cycles from the decision space, we propose the *token-split* method. If a pair of nodes (v_i, v_j) is selected to form a route, a token of one unit, $t_{ij} + t_{ji} = 1$, is unconditionally split between them such that v_i and v_j receive t_{ij} and t_{ji} respectively. Note that the only constraint on t_{ij}, t_{ji} is that they are nonnegative reals. However, to exclude the above-stated cycles, we impose that each node must receive a total amount of tokens that is *strictly less* than 1. To see how this works, suppose that we have a cycle of m nodes and m links selected. Then, the total tokens to be split among them equals exactly m . In this case it is impossible to find a way to split tokens between consecutive pairs in the loop such that each node receive *strictly less* than 1. To formulate such strict inequality by a non-strict one, a threshold parameter $0 < \epsilon < \frac{1}{2}$ may be used. Then, the no-loop constraint can be stated as:

$$\forall v_n \in \mathcal{V}, \sum_{e_{ij} \in \mathcal{E} | i=n} t_{ij} \leq 1 - \epsilon$$

However, we must also be sure that such restriction does not rule out arbitrary paths in the solution space. Suppose that we have a path of $m \geq 3$ nodes with $m - 1$ links selected. Then, we show that it is possible to split the total $m - 1$ tokens respecting the previous constraints if $\epsilon \leq \frac{1}{m}$. To prove this, we can split the tokens such that the first $m - 1$ nodes receive exactly $1 - \epsilon$ token each, and as a consequence, the last one receives $(m - 1) - (m - 1)(1 - \epsilon)$ or $(m - 1)\epsilon$ token. This is explained graphically as follows:



To respect the no-cycle condition at the last node, we have $(m - 1)\epsilon \leq 1 - \epsilon$ which implies $\epsilon \leq \frac{1}{m}$ which completes the proof. To sum it all, if we set the parameter $\epsilon = \frac{1}{|\mathcal{V}|}$, where $|\mathcal{V}|$ is the total number of nodes, then all possible loops are excluded from the solution space without excluding any possible (simple) path from a source to a destination. Formally,

the no-cycle constraints are given by:

$$\forall e_{ij} \in \mathcal{E}, x_{ij} + x_{ji} \leq 1 \quad (6)$$

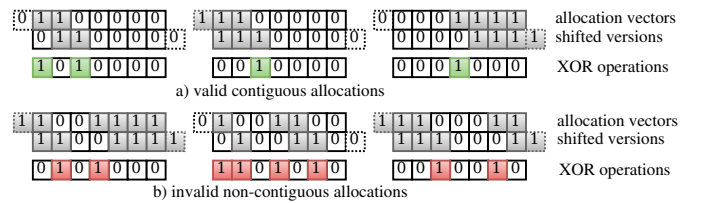
$$\forall e_{ij} \in \mathcal{E}, t_{ij} + t_{ji} = x_{ij} + x_{ji} \quad (7)$$

$$\forall v_n \in \mathcal{V}, \sum_{e_{ij} \in \mathcal{E} | i=n} t_{ij} \leq 1 - \frac{1}{|\mathcal{V}|} \quad (8)$$

Given the half-duplex mode hardware constraint in D2D, UEs cannot simultaneously transmit and receive on SL. Therefore, active nodes must switch back and forth between roles. In order to reduce the end-to-end delay, we require all non-successive nodes in a path to transmit in one period while their respective partners are listening to them, and on the next period, they swap roles. This principle of operation forces the links along a path to be scheduled in alternating manner. The net effect of these assumptions is that the SL scheduler switches every SL frame between two sets of active UEs in order to maintain the ongoing flows. In other words, the active nodes $\mathcal{V}_H \subseteq \mathcal{V}$ are divided into two sets: $\mathcal{V}_H^0 = \{v_n \in \mathcal{V} \mid H_n = 0\}$ and $\mathcal{V}_H^1 = \{v_n \in \mathcal{V} \mid H_n = 1\}$ where H_n are binary variables attached to the nodes. Hence a pair of nodes, having an active link between them, cannot be in the same *half-duplex set* (period):

$$\forall e_{ij} \in \mathcal{E}, x_{ij} \leq H_i + H_j \leq 2 - x_{ij} \quad (9)$$

In addition to its assigned half-duplex period, a transmitting node needs also frequency resources. In line with LTE-D2D standard, we assume a Frequency Division Duplex (FDD) cellular network where we assign a bandwidth B_{SL} , composed of Ω contiguous OFDMA RBs, to the SL operation. Note that only contiguous RB allocations are feasible within this bandwidth because the SL has the same communication proprieties as the UL [9]. We represent the allocated RBs for a node v_n , by a vector of 0-1 variables R_n^ω for $\omega = 1, 2, \dots, \Omega$, where the variable R_n^ω indicates whether the RB number ω is allocated to v_n . To formulate the contiguity constraints, we use the *Hamming distance*. The Hamming distance $d_H(V_1, V_2)$ between the vectors V_1 and V_2 is the number of positions at which the two vectors differ. The Hamming distance between two 0-1 vectors, $V_1 = [V_1^1, V_1^2, \dots, V_1^\Omega]^T$ and $V_2 = [V_2^1, V_2^2, \dots, V_2^\Omega]^T$, is the sum of component-wise XOR operation between the vectors (i.e. $d_H(V_1, V_2) = \sum_{i=1}^{\Omega} V_1^i \oplus V_2^i$). To check an allocation vector $[R_n^1, R_n^2, \dots, R_n^\Omega]^T$ for contiguity, we remark that the Hamming distance between $[R_n^1, R_n^2, \dots, R_n^{\Omega-1}]^T$ and its shifted version $[R_n^2, R_n^3, \dots, R_n^\Omega]^T$ is less than 2 if the $R_n^1 = 0$ and is less than 1 if $R_n^1 = 1$ as illustrated as follows:



Formally, this constraint is express as:

$$\forall v_n \in \mathcal{V}, \sum_{\omega=1}^{\Omega-1} R_n^\omega \oplus R_n^{\omega+1} \leq 2 - R_n^1 \quad (10)$$

In addition, RBs are allocated for some node v_n only when the node is a transmitter (i.e. one of its outgoing links is selected). Formally:

$$\forall v_n \in \mathcal{V}, \sum_{\omega=1}^{\Omega} R_n^\omega \leq \sum_{e_{ij} \in \mathcal{E} | i=n} x_{ij} \quad (11)$$

To increase the reutilization of RBs and to reduce power

consumption, we require that nodes are not allocated RBs beyond the request of the associated flow or formally:

$$\forall f^k \in \mathcal{F}, \forall v_n \in \mathcal{V}, \sum_{\omega=1}^{\Omega} R_n^{\omega} \leq \Omega + (D^k - \Omega) A_n^k \quad (12)$$

Note that the relation between flow bit-rate R^k and the respective demand for RBs D^k is defined by [9] as:

$$R^k = \frac{\text{TBS}(\text{MCS}, D^k)}{C} [\text{mbps}] \quad (13)$$

where TBS is the MAC transport block size function in bits as defined in [9] considering a baseline *Modulation and Coding Scheme* (MCS) for the SL. C is a constant equal to 1000.

In face of the reutilization of RBs, system performance is limited by interference caused by nodes transmitting using the same RB. To deal with interference, we assume a fixed power density scheme for the D2D emission. According to this scheme, the total emission power $S_{\text{tx},n}$ of a node is proportional to the number of allocated RBs $\sum_{\omega=1}^{\Omega} R_n^{\omega}$. Formally,

$$S_{\text{tx},n} = \Psi_{t,n} \cdot \sum_{\omega=1}^{\Omega} R_n^{\omega} \quad [\text{mW}] \quad (14)$$

Furthermore, we assume a common emission power density, $\Psi_{t,n}$ [mW/RB], for all the D2D nodes (i.e., $\forall v_n \in \mathcal{V}, \Psi_{t,n} = \Psi_t$). Following the same *per-RB* treatment and assuming *flat block-fading* channel model, the overall Signal-to-Interference-plus-Noise Ratio (SINR) on the link e_{ij} is equal to:

$$\gamma_{ij} = \frac{g_{ij} \Psi_{t,i}}{\sum_{v_n \in \mathcal{V}} g_{nj} \Psi_{t,n} + \Psi_{\sigma}} \quad (15)$$

where Ψ_{σ} and $\Psi_{t,n}$ represent the spectral densities (per RB) of the thermal noise and the transmission from v_n , and g_{ij} is the channel gain between the node pair (v_i, v_j) .

Furthermore, additional variables $R_n^{\omega,0}$ and $R_n^{\omega,1}$ are defined to indicate whether the RB ω is used by v_n in $\mathcal{V}_H^0$ or $\mathcal{V}_H^1$ (i.e. half-duplex set of frames), respectively. Formally,

$$\forall (v_n, \omega) \in \mathcal{V} \times [1, \Omega], R_n^{\omega,0} \triangleq R_n^{\omega} - R_n^{\omega,1} \quad (16)$$

$$\forall (v_n, \omega) \in \mathcal{V} \times [1, \Omega], R_n^{\omega,1} \triangleq H_n \cdot R_n^{\omega} \quad (17)$$

An additional set of link-level auxiliary 0-1 variables are introduced as follows:

$$\forall (e_{ij}, \omega, p) \in \mathcal{E} \times [1, \Omega] \times \{0, 1\}, R_{ij}^{\omega,p} \triangleq R_i^{\omega,p} \quad (18)$$

$$\forall (e_{ij}, \omega, p) \in \mathcal{E} \times [1, \Omega] \times \{0, 1\}, \phi_{n,ij}^{\omega,p} \triangleq R_n^{\omega,p} \cdot R_{ij}^{\omega,p} \quad (19)$$

where $R_{ij}^{\omega,p}$ indicates if the RB ω is used for the scheduled link e_{ij} during the p^{th} half duplex set, $\phi_{n,ij}^{\omega,p}$ is an interference indicator between node v_n and link e_{ij} on the RB ω .

To adhere to a linear formulation, further steps are needed to linearize the XOR-terms in Constraint (10) and the product terms in Constraints (17), (18) and (19).

We make use of a standard technique to linearize each XOR-term $x \oplus y$ by introducing an additional auxiliary 0-1 variable $\lambda_{xy}^{\oplus}$ add four more linear constraints as follows:

$$(\lambda_{xy}^{\oplus} \geq x - y), (\lambda_{xy}^{\oplus} \geq y - x), (\lambda_{xy}^{\oplus} \leq x + y), (\lambda_{xy}^{\oplus} \leq 2 - x - y) \quad (20)$$

We use another standard technique to linearize each product term $x \cdot y$ by introducing an additional auxiliary 0-1 variable $\lambda_{xy}^{\odot}$ add four more linear constraints as follows:

$$(\lambda_{xy}^{\odot} \leq x), (\lambda_{xy}^{\odot} \leq y), (\lambda_{xy}^{\odot} \geq x + y - 1) \quad (21)$$

To optimize the performances by minimizing interferences, SINR must be upper-bounded by a common threshold γ . To formulate this constraint on RB allocations, we translate this limit (i.e., $\text{SINR} \leq \gamma$) into the inequality $\mathcal{N} + \mathcal{I} \leq P_r / \gamma$ where P_r is the received power.

Algorithm 1 JRW-D2D pseudo-code

```

1: for each SL frame  $\tau$  do
2:   for each  $f^k \in \mathcal{F}_A$  do ▷ Arriving flows
3:     if  $s_k$  and  $d_k \in$  the same component of  $\mathbb{G}$  then
4:        $\mathcal{F}_W \leftarrow \mathcal{F}_W \cup \{f^k\}$ 
5:     end if
6:   end for
7:   for each  $f^k \in \mathcal{F}_{\text{FIN}}$  do ▷ Finished flows
8:      $\mathcal{V}_D \leftarrow \mathcal{V}_D \cup \text{NodesOF}(f^k)$ 
9:   end for
10:  Construct the MILP model as in formula (23)
11:  Solve the MILP model using Algorithm 2
12:  for each  $f^k \in \mathcal{F}_W$  do
13:    if  $A_{s_k}^k = 1$  then ▷ Flow is admitted
14:      Configure the path according to  $\mathbb{P}^k$ 
15:    end if
16:  end for
17:   $p \leftarrow \tau \bmod 2$ 
18:  for each  $v_n \in \mathcal{V}_H^p$  do
19:    Allocate RBs according to  $[R_n^1, R_n^2, \dots, R_n^{\Omega}]^T$ 
20:  end for
21: end for

```

$$\forall (e_{ij}, \omega, p) \in \mathcal{E} \times [1, \Omega] \times \{0, 1\}, \Psi_{\sigma} R_{ij}^{\omega,p} + \sum_{n \neq i} g_{nj} \Psi_t \cdot \phi_{n,ij}^{\omega,p} \leq \frac{g_{ij} \Psi_t}{\gamma} R_{ij}^{\omega,p} \quad (22)$$

As stated before, the function of our eNodeB is to schedule the SL resources in order to support the ongoing (already-admitted) flows and to handle newly-arriving flows trying to admit some of them when possible. In doing so, the objective is to maximize the overall utilization of system resources (nodes and RBs) while serving the maximum possible number of flows. To reach such objective, our utility function can be decomposed into three goals: 1) maximizing the total number of allocated RBs, 2) maximizing the number of admitted flows, and 3) minimizing the total hop-count of the reserved paths.

We propose to formulate these goals as single objective-function of weighted-sums to complete the MILP formulation, developed so far, as follows:

$$\max. \quad \alpha_B \sum_{\substack{x_{ij}, A_{n,t_{ij}}^k \\ H_n, R_n^{\omega}, \dots}} \sum_{\substack{v_n \in \mathcal{V} \\ \omega \in [1, \Omega]}} R_n^{\omega} + \alpha_A \sum_{f^k \in \mathcal{F}} A_{s_k}^k - \alpha_N \sum_{\substack{v_n \in \mathcal{V} \\ f^k \in \mathcal{F}}} \sum A_n^k$$

subject to:

$$(2) \text{ to } (12), (16) \text{ to } (19) \text{ and } (22)$$

$$t_{ij} \in [0, 1] \subset \mathbb{R}, \text{ all other variables } \in \{0, 1\} \quad (23)$$

where the normalizing factors defined by:

$$\alpha_B = \frac{1}{\Omega |\mathcal{V}|}, \alpha_A = \frac{1}{|\mathcal{F}|}, \alpha_N = \frac{1}{|\mathcal{V}|} \quad (24)$$

V. PROPOSAL: JRW-D2D

In this paper, we propose novel strategy named **Joint Routing and Wireless allocation in D2D communications** (JRW-D2D) to solve the optimization problem described above. Our proposal is based on *Branch-and-Cut* algorithm [4]. The latter is well-known optimization algorithm and efficient to solve the general class of *Mixed-Integer-Linear-Programming* (MILP) problems. JRW-D2D proceeds as follow. First, the binary variables are relaxed by allowing them to admit continuous values between 0 and 1. Then, the relaxed problem is solved by *simplex* algorithm. If the latter converges to an optimal solution with at least fractional value for a variable, then a *branch* is introduced on that variable. A branch means that two subproblem nodes are

Algorithm 2 MILP resolution**Input:** MILP Model P^0 as defined in formula (23)**Output:** Solution value for V^* as $[x_{ij}, A_n^k, t_{ij}, H_n, R_n^\omega, \dots]$

```

1: Push the initial problem  $P^0$  onto the stack  $\mathcal{S}$ 
2:  $f^* \leftarrow -\infty$  ▷ Initial value for Objective function
3:  $I \leftarrow 0$  ▷ Counter
4: while  $\mathcal{S} \neq \emptyset \wedge I \leq I_{\max}$  do
5:    $I \leftarrow I + 1$ 
6:   Pop a problem from  $\mathcal{S}$  as  $P$ 
7:   Let  $\tilde{P}$  be the relaxed form of  $P$  with continuous  $V^*$ 
8:   Solve  $\tilde{P}$  using simplex yielding  $\tilde{V}$  and  $\tilde{f}$ 
9:   if not feasible or  $\tilde{f} \leq f^*$  then go to 17
10:  if  $\tilde{V}$  are all 0 or 1 except for  $t_{ij}$  then
11:     $V^* \leftarrow \tilde{V}$ ,  $f^* \leftarrow \tilde{f}$  and go to 17
12:  else
13:    Choose the closest variable to 0.5 as  $v^\dagger$ 
14:    Add a cut  $v^\dagger \leq 0$  to  $P$  and push it onto  $\mathcal{S}$ 
15:    Add a cut  $v^\dagger \geq 1$  to  $P$  and push it onto  $\mathcal{S}$ 
16:  end if
17: end while
18: return the solution value  $V^*$ 

```

scheduled to be solved recursively with additional *cuts* (i.e., additional inequality constraints). Each cut bounds the variable in subproblems by 0 or 1. Each subproblem is, in its turn, relaxed again and the whole process repeats until finding a set of feasible integral solutions that includes the optimal one. However, a scheduled problem node is *pruned* if its objective-function value in the relaxed solution is worse than the best integral solution found so far. Pruning a node means that the latter cannot generate further subproblems. Hence, extensive search for optimal integral solution is avoided. Algorithms 1 and 2 illustrate the pseudo-code of our proposal JRW-D2D. It should be noted that we also introduce a bound on the number of recursive iterations to limit the execution time.

VI. PERFORMANCE EVALUATION

In this section, we will gauge the performance of our proposal JRW-D2D based on extensive simulations. First of all, we will briefly describe the network simulation environment NS-3 which we augmented to support LTE-D2D protocol stack. Then, we will detail the studied scenario in this paper. Afterwards, we will define the performance metrics. Finally, we will analyze the simulation results and discuss the effectiveness of our proposal.

A. Network simulation environment

The NS-3 software package [10], which is written in C++, provides powerful open-source tools to implement a wide variety of network simulation scenarios and application using different degree of abstractions and reference technologies. NS-3 provides substantial support for a variety conventional 3GPP LTE simulation scenarios through the module NS-3/LTE. Unfortunately, the latter does not support the LTE-D2D standard. To the best of our knowledge, this is the case for all available network simulators in this respect. This is, in part, due to the fact that LTE-D2D is a relatively new standard. To achieve our goal, we extended the NS-3/LTE modules to include the necessary LTE-D2D protocol stack. We developed the PHY, MAC and PDCP/RLC procedures along with the signaling between the eNodeB and UEs. The signaling is necessary to i) configure the SL parameters, ii) establish the SL radio bearers (SLRB), and iii) exchange the SL reports and grants.

TABLE I
SIMULATION PARAMETERS

Parameter	Value
Cell Radius R_{cell}	1 km
UL/SL Frequency f_{UL}	1930 MHz
UL/SL (Reference) Bandwidth B_{UL}	5 MHz (25 LTE RBs)
SL RBs Used Actually Ω	14 LTE RBs
SL frame (LTE-D2D SC-Period)	40 subframes (40 ms)
Data Part in SL frame	32 subframes
UE SL Power Transmit Density Ψ_t	-4 dBm/RB
Noise Spectral Density Ψ_n	-121.45 dBm/RB
LTE MCS Index used in SL	9 (QPSK)
UE Density λ_{UE}	{10, 15, 20, 25, 30, 35, 40} per km ²
UE-UE SNR Threshold γ_{TOPO}	10 dB
Scheduling SINR Threshold γ	6 dB
Flow Simulation Period	10 seconds
Flow Arrival Process	Poisson Process
Flow Arrival Rates λ_{FL}	{10, 20} flows/second
Flow Duration Random Variable	Exponential
Flow Duration Mean λ_{DUR}	1 second
Flow Bit Rate Classes	{25, 50, 75, 100, 125, 150, 175, 200} kbps

B. Network simulation setup

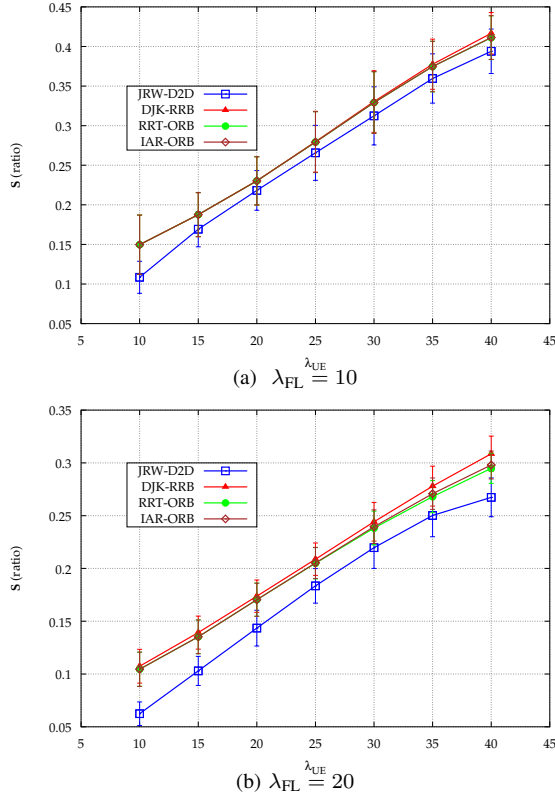
In line with our formulation in section IV, we run simulations for a network composed of one macro-cell LTE-A with radius $R_{\text{cell}} = 1$ km. The geographical deployment of UEs inside the cell follows a Poisson Point Process distribution with a density λ_{UE} nodes per km² for values in the set {10, 15, 20, 25, 30, 35, 40}. The LTE-A macro-cell is configured to work in FDD mode with an UL frequency of 1930 MHz and a bandwidth of 5 MHz (i.e., 25 RBs). The eNodeB configures SL bandwidth to share the same as UL. However, The eNodeB allocates scheduled-resources pool only $\Omega = 14$ RBs for the offloading operation over SL. UEs are configured to transmit on SL with a common power density of $\Psi_t = -4$ dBm/RB (i.e., maximum of 10 dBm over the whole 5 MHz). To model the SL path-loss, we use the WINNER II B2-LOS channel model [11]. The SL-Period (SL frame) is configured to be 40 milliseconds (i.e., 40 subframes), which is the minimum possible value in the standard, of which 32 subframes is used for the data transmission. The eNodeB, using a SNR reports, builds the D2D network topology. A link is considered part of the network if the respective SNR is greater than $\rho_{\text{TOPO}} = 10$ dB. Traffic flows are generated according to a Poisson process with an arrival rate $\lambda_{\text{FL}} \in \{10, 20\}$ flows per second. On the other hand, each flow is assumed to have a Constant Bit-Rate (CBR) traffic randomly selected from predefined CBR classes. Flow duration distribution is simulated to follow an exponential random variable with a mean duration of $\lambda_{\text{DUR}} = 1$ second. Sources and destinations are chosen from a random uniform distribution. TABLE I summarizes the main parameters used in our network simulation. For the evaluation of results, the confidence level is set to 95%.

C. Performance metrics

We define the following metrics to evaluate our proposal:

- 1) $\mathbb{S}$ is the ratio of the flows offloaded by the D2D subnetwork.
- 2) $\mathbb{A}$ is the maximum number of scheduled flows simultaneously.
- 3) $\mathbb{L}$ is the average of flow packet loss in each simulation run.
- 4) $\mathbb{H}$ is the average number of hops in the offloading path in each simulation run.

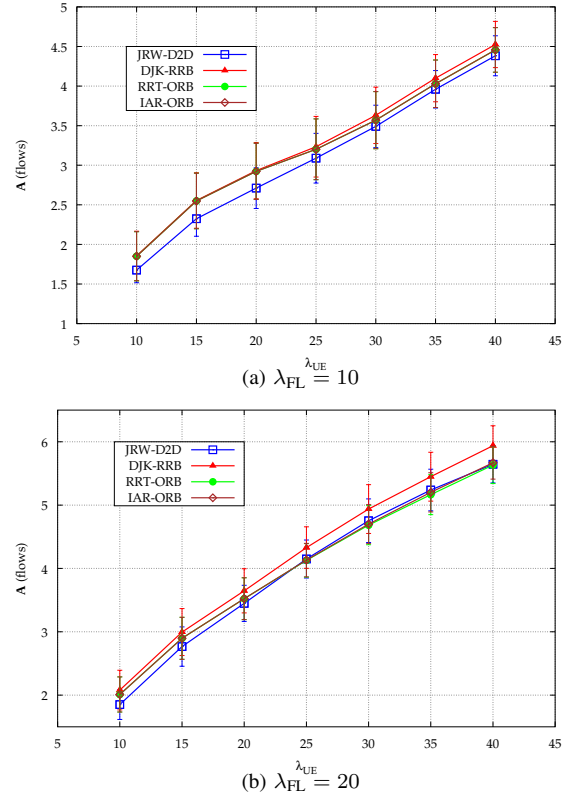
In addition we compare the performance of our proposal JRW-D2D with the following alternative routing and resource strategies:

Fig. 4. $\mathbb{S}$ w.r.t the nodes density λ_{UE} .

- 1) DJK-RRB: is a pure path strategy that aims to find the optimal routing trees using Dijkstra algorithm and then, allocates RB randomly.
- 2) RRT-ORB: is a pure resource block oriented strategy that finds the routing trees randomly using random walk on the topology graph, and allocates RB optimally.
- 3) IAR-ORB: is a heuristic scheme composed of interference aware routing based on Dijkstra algorithm. In this variant, the link costs to minimize are the total interference level on the link taking into consideration the actual state of network before accepting the new flows. Then, the resource block allocations are done optimally.

D. Simulation results

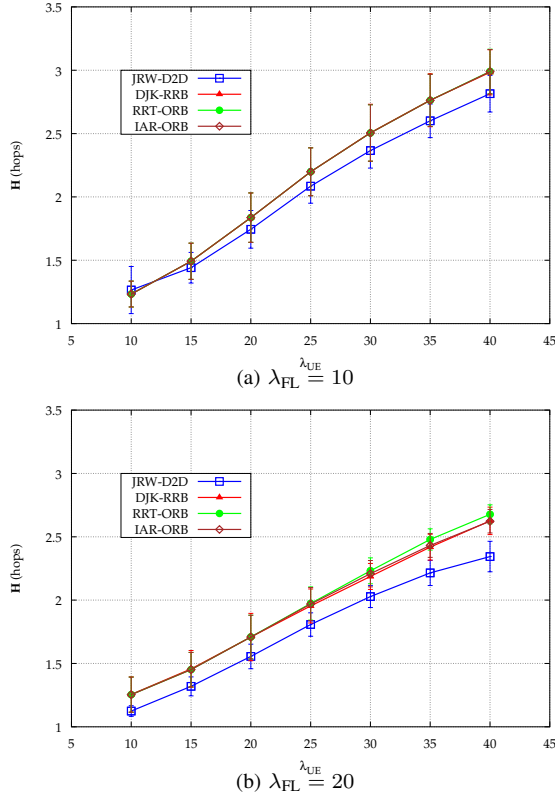
First, we evaluate our algorithm JRW-D2D regarding its offloading capability and in comparison with the alternative strategies DJK-RRB, RRT-ORB, IAR-RRB. To do so, we calculate the ratio of the flows offloaded over D2D subnetwork. Fig. 4 illustrates $\mathbb{S}$ with respect to the density of UEs and under two traffic conditions $\lambda_{FL} = 10$ and $\lambda_{FL} = 20$ flows per second. For small values of λ_{UE} , it is straightforward to see that $\mathbb{S}$ increases in proportion to λ_{UE} . This means that as D2D node density increases, more flows will succeed to be routed through the D2D network. This is expected because, when the density increases, the probability of forming reliable D2D links rises accordingly. And as a consequence, the D2D network capacity to absorb random flows also grows. We remark that DJK-RRB outperforms the other schemes in general. This is expected since DJK-RRB routes flows over the fewest possible nodes. As a result, it allows for more flows to be admitted into the network. Taking DJK-RRB as a baseline, we note that our proposal JRW-DD has a flow acceptance rate of $\Delta\mathbb{S} = 1\%$ less than the baseline DJK-RRB, in average, for $\lambda_{FL} = 10$ as depicted in Fig. 4a. On the other hand, Fig. 4b shows the

Fig. 5. $\mathbb{A}$ w.r.t node density λ_{UE} .

situation under more traffic pressure, $\lambda_{FL} = 10$, where the acceptance rate drops for all schemes while the performance gap of JRW-DD increases to be around $\Delta\mathbb{S} = 3\%$ in average with respect the leader DJK-RRB.

To complement the evaluation of the offloading capability, we measure the degree concurrency in utilizing the D2D subnetwork. To this end, we measure the average of the maximum number of flows offloaded simultaneously over D2D subnetwork. This measure is conveyed by the metric $\mathbb{A}$, shown in Fig. 5, which demonstrates to what degree the different scheme are successful to utilize system resources concurrently. In a manner consistent the evolution of $\mathbb{S}$, the evolution of $\mathbb{A}$ is depicted in Fig. 5 under the two traffic conditions $\lambda_{FL} = 10$ and $\lambda_{FL} = 20$. We note that the metric $\mathbb{A}$ increases in response to an increase in λ_{UE} . This reflect the fact that, in a denser topology, more nodes are available in the network to route concurrent flows circumventing the restriction due to maximum one flow per node. Again, we note that DJK-RRB is the leader of the group where the our scheme JRW-DD was able to offload slightly fewer simultaneous flows than the others with performances gaps $\Delta\mathbb{A} \leq 0.25$ simultaneous flows as indicated in Fig. 5a and Fig. 5b.

To quantify the QoS presented to the offloaded flows we focus on latency and packet loss rate. Fig. 6 illustrates the performance in terms of $\mathbb{H}$ metric which count the number of hops in the routing paths. This metric indicates the QoS presented to flows in terms of latencies where shorter is better. Specifically, the end-to-end and the average packet delays are proportional to $\mathbb{H} \times T_{SL}$. Fig. 6a and Fig. 6b point out that the average number of hops increases almost linearly in accordance with the density of nodes λ_{UE} . The figure reveals that JRW-D2D leads to shorter paths in average than the others. This seems paradoxical in particular when comparing

Fig. 6. $\mathbb{H}$ w.r.t node density λ_{UE}

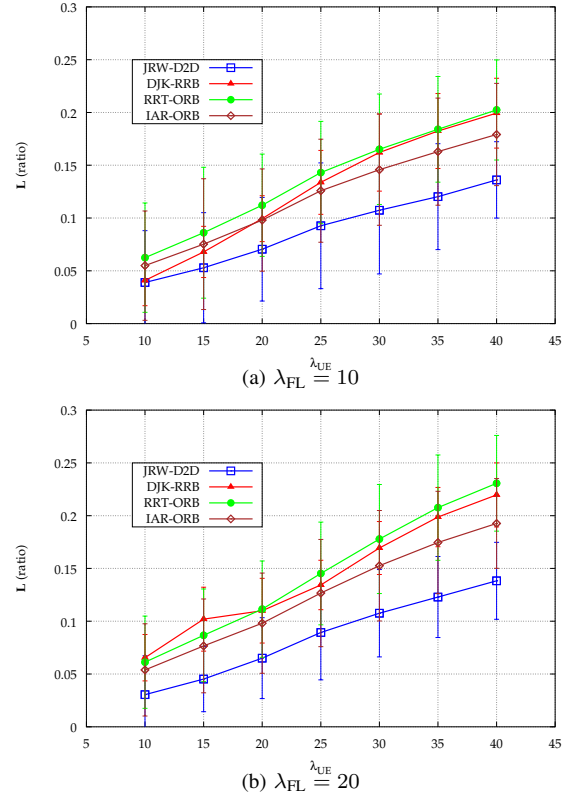
to DJK-RRB. However lower values of $\mathbb{H}$ are an artifact of JRW-D2D being biased to accept flows with shorter paths at the expense of blocking some long path flows.

Moreover, to quantify the QoS in terms of the packet error rate at the IP level, Fig. 7 illustrates the average packet loss ($\mathbb{L}$) in flows as a function of the UEs' density for $\lambda_{FL} = 10$ and $\lambda_{FL} = 20$ conditions respectively. In Fig. 7, it is straightforward to see that our scheme JRW-D2D outperforms the other schemes thanks to their capability to take into consideration interference in OFDMA RB blocks allocation. However, RRT-ORB performs badly in general, which may seem paradoxical. It is straightforward to see that such a behavior will lead to higher transmission delays. Being interference-aware in routing and resource allocation makes JRW-D2D more robust against the packet loss. In fact, the latter succeeds to maintain $\mathbb{L}$ below 0.13 and 0.14 for both traffic conditions $\lambda_{FL} = 10$ and $\lambda_{FL} = 20$ flow per seconds respectively as depicted in Fig. 7a and Fig. 7b.

In summary, network simulations show that JRW-D2D outperforms the variants in terms of reliability at the expense of small performance gaps with respect to latency and offloading capacity.

VII. CONCLUSION

In this paper, we addressed the problem of joint routing and OFDMA resource allocation in LTE-D2D multihop networks, considering LTE-D2D-specific constraints, namely, the half-duplex operation and the contiguity in RB allocations. An offloading application, as a use case, was conceived where data from the eligible flows are routed over the D2D multihop network and the eNodeB host the control plane. We proposed a MILP formulation for the problem and a novel scheme named JRW-D2D based on the branch-and-cut algorithm. We validated our proposal with simulating the whole LTE D2D

Fig. 7. $\mathbb{L}$ w.r.t node density λ_{UE} .

protocol stack in the NS-3 network simulator. The results obtained are very satisfying in terms of optimality, ratio of admitted D2D flows and latency in comparison to other implemented basic one-sided optimal strategies and another interference-aware heuristic scheme.

REFERENCES

- [1] Cisco Visual Networking, "Cisco Global Cloud Index: Forecast and Methodology, 2015-2020," *White paper*, 2016.
- [2] A. Asadi, Q. Wang, and V. Mancuso, "A Survey on Device-to-Device Communication in Cellular Networks," *IEEE Communications Surveys and Tutorials*, vol. 16, no. 4, pp. 1801–1819, Fourthquarter 2014.
- [3] P. Gandotra and R. K. Jha, "Device-to-Device Communication in Cellular Networks: A Survey," *Journal of Network and Computer Applications*, vol. 71, pp. 99–117, 2016.
- [4] J. E. Mitchell, "Integer Programming: Branch and Cut Algorithms," in *Encyclopedia of Optimization*, C. A. Floudas and P. M. Pardalos, Eds. Springer, 2009, pp. 1643–1650.
- [5] L. Babun, A. İ. Yürekli, and İ. Güvenç, "Multi-hop and D2D communications for extending coverage in public safety scenarios," in *2015 IEEE 40th Local Computer Networks Conference Workshops (LCN Workshops)*, Oct 2015, pp. 912–919.
- [6] V. Bhardwaj and C. R. Murthy, "On optimal routing and power allocation for D2D communications," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 3063–3067.
- [7] J. Li, W. Xia, S. Xing, and L. Shen, "Transmission scheduling and congestion control for multi-hop D2D underlying cellular networks," in *2014 IEEE 25th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC)*, Sept 2014, pp. 996–1000.
- [8] 3GPP, "Proximity-based services (ProSe); Stage 2," 3rd Generation Partnership Project (3GPP), TS 23.303, Dec. 2016. [Online]. Available: <http://www.3gpp.org/DynaReport/23303.htm>
- [9] —, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures (Release 14)," 3rd Generation Partnership Project (3GPP), TS 36.213, Sep. 2017, v14.4.0. [Online]. Available: <http://www.3gpp.org/DynaReport/36213.htm>
- [10] G. F. Riley and T. R. Henderson, "The ns-3 Network Simulator Modeling and Tools for Network Simulation," in *Modeling and Tools for Network Simulation*, K. Wehrle, M. Güneş, and J. Gross, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, ch. 2, pp. 15–34.
- [11] Y. d. J. Bultitude and T. Rautiainen, "IST-4-027756 WINNER II D1.1.2 V1.2 WINNER II Channel Models," 2007.