Throughput maximization in multi-slice cooperative NOMA-based system with underlay D2D communications

Asmaa Amer, Sahar Hoteit, Jalel Ben-Othman

Throughput Maximization in Multi-Slice Cooperative NOMA-based System with underlay D2D Communications

Asmaa Amer[a,*], Sahar Hoteit[a], Jalel Ben Othman[a,b]

[a] *Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes, Gif-sur-Yvette, 91190, France*
[b] *Université Sorbonne Paris Nord, Paris, France*

Abstract

The fifth generation (5G) and beyond-5G networks aim to meet the rapidly growing traffic demands while considering the scarcity of radio resources and the heterogeneity of services and technical requirements. Non-Orthogonal Multiple Access (NOMA) has been considered a key technology to address resource scarcity by enabling more users to share the resources. Furthermore, network slicing tackles the requirements' heterogeneity by partitioning the physical network into multiple logical slices. In this study, the aforementioned key technologies are adopted to maximize the overall throughput and satisfy the technical requirements of a multi-slice system. We formulate an optimization problem in a multi-slice cooperative NOMA-based system with underlay D2D communications that jointly addresses user grouping, radio resource blocks allocation, and D2D admission. Its objective is to maximize the overall system throughput while considering each slice's constraints. Given the complexity of the optimization problem, we propose a three-step, low-complexity solution: a matching theory-based approach for the user grouping and the D2D admission sub-problems, and a heuristic approach for the resource blocks allocation sub-problem. Numerical results demonstrate the:

*This paper was partially presented at IEEE International Conference on Communications, ICC, Rome, Italy, May 2023 [1].

*Corresponding author

Email addresses: asmaa.amer@centralesupelec.fr (Asmaa Amer), sahar.hoteit@centralesupelec.fr (Sahar Hoteit), jalel.benothman@centralesupelec.fr (Jalel Ben Othman)

1) low complexity of the proposed solution; 2) high impact of interference cancellation imperfection; 3) superior performance of the proposed solution compared to literature baselines. Specifically, under dense network, our solution achieves up to 34% enhancement in overall system throughput, up to 35% improvement in D2D admission, and up to 125% and 32%, respectively, in cellular users and D2D pairs satisfaction. It also outperforms under strict eMBB and URLLC requirements. Also, results show significant overestimation in Shannon's evaluation of URLLC throughputs considered in some papers from the literature compared to the finite block length evaluation considered in our work, particularly at higher URLLC reliability requirements.

Keywords: Cooperative non-orthogonal multiple access (CNOMA), Device-to-device (D2D) communications, eMBB, Resource allocation, URLLC

---

1. Introduction

The rapid development of wireless technology has led to a substantial increase in the number of connected devices and the resulting data traffic. Cisco expects an increase in the number of networked devices to reach 29.3 billion by the end of 2023 [2, 3]. Furthermore, these devices reflect diverse applications and consequently have heterogeneous services and technical requirements in terms of throughput, latency, reliability, massive connectivity, and energy consumption. This heterogeneity presents a challenge to the one-size-fits-all architecture of the previous network generations as they are unable to completely handle it [4, 5]. Three main service categories have been considered as the fifth generation (5G) pillars, they are as follows: *(i)* enhanced mobile broadband (eMBB), which refers to services requiring high data rates; *(ii)* ultra-reliable and low latency communications (URLLC) which refers to mission-critical applications with high reliability and low latency; *(iii)* massive machine-type communications (mMTC), which stands for massively connected and energy-constrained services [4]. Next-generation networks (i.e., beyond 5G (B5G)/sixth generation (6G)) are envisioned to extend upon these 5G basic service categories into services with combined features and more stringent requirements, including especially much higher data rates, and sub-millisecond latency [6, 7].

In light of these stringent requirements, new key technologies are considered as promising solutions for next-generation networks; these technologies

include the evolution of certain 5G technologies as well [8]. Among these key technologies, Non-Orthogonal Multiple Access (NOMA) has been reported by the International Telecommunication Union (ITU) as one of the future technology trends towards 2030 and beyond [9]. NOMA has been considered as a viable solution capable of boosting massive connectivity [10, 11] and achieving spectral [12] and energy efficiency [13], and lower latency [14, 15] for 5G and B5G systems. These benefits stem from NOMA's ability to efficiently manage multiple user access, a crucial factor in overcoming the resources scarcity challenge. This is particularly in light of the ten-fold increase in connection density envisaged in next-generation networks compared to 5G networks [8]. In the previous network generations, Orthogonal Multiple Access (OMA) technique has been utilized. It consists of allocating resource blocks (RBs) to users in an orthogonal manner [16]. OMA technique avoids interference between users but fails to meet the massive connectivity demand of 5G and B5G networks. Unlike OMA, NOMA[1] technique permits multiple users to share the same RBs simultaneously; these users form a NOMA group [17]. NOMA employs superposition coding (SC) at the transmitter side and successive interference cancellation (SIC) at the receiver side. In other terms, in downlink transmission, the base station (BS) transmits a superposed signal (i.e., a linear combination of the messages of the users within the same NOMA group) where each user is differentiated by a power coefficient (i.e., power-domain multiplexing). Then, to suppress a portion of the resulting interference on the users' side, each user performs the SIC process whereby rather than considering the messages of other users as noise, it decodes and eliminates the messages of all users having lower channel gains to BS one by one before decoding its own [18]. However, the messages of users with higher channel gain are not suppressed by SIC; they instead act as interference known as co-channel interference.

Besides the conventional NOMA, the cooperative NOMA (CNOMA) scenario has been proposed in [19]. CNOMA consists of employing cooperative communications to extend the coverage, and enhance the performance of the weak or cell-edge users. Taking advantage of SIC, the strong or cell-center user acts as a relay and forwards to the weaker user in the same NOMA group, its message. A significant part of NOMA-based systems literature is

[1]In this paper, NOMA refers to power-domain NOMA; code-domain NOMA is out of the scope of this paper.

devoted to CNOMA, but the frequency resource block allocation problem in CNOMA systems is not thoroughly investigated.

Likewise NOMA, Device-to-Device (D2D) communication has been considered as a promising key technology for proximity communication. D2D communications provide diverse proximity-based services by enabling direct communication between proximate devices (i.e., D2D pairs), bypassing the BS [20]. Thus, D2D communication can greatly enhance the system throughput, and reduce the end-to-end latency. As per CISCO's annual internet report (2018-2023), the D2D links, referred to as machine-to-machine (M2M) links, are expected to increase up to 14.7 billion by the end of 2023 [2, 21]. The operation of these D2D communication links can be categorized in terms of spectrum utilization into two modes, namely, in-band and out-band modes. In the in-band mode, the D2D devices can utilize the licensed spectrum (i.e., the cellular spectrum). The in-band mode can be further categorized into: *(i)* in-band underlay mode, i.e., the cellular users and D2D devices share the same resource blocks at the expense of cellular-D2D interference, and *(ii)* in-band overlay mode, i.e., a number of resources are reserved for D2D communications only, at the expense of wasting resources at the time of low D2D demand. However, in the out-band mode, D2D communication utilizes the unlicensed spectrum (i.e., the Industrial, Scientific, Medical (ISM) band). Thus, D2D communication will be vulnerable to uncontrolled interference from devices and technologies operating in the crowded ISM band, and consequently, their services are hard to be satisfied [22]. Overall, the in-band underlay mode, where controlled interference management can be applied, provides the higher spectral efficiency[2]. Motivated by the ultra-massive connectivity demand of the expected dense next-generation networks, incorporating the underlay D2D communications in a NOMA-based cellular system reflects a dense and diverse practical B5G/6G scenario. Our paper is focusing on this scenario. However, the challenge in such a scenario refers to effectively managing the interference that results from sharing cellular users' resources with D2D devices while ensuring the satisfaction of their heterogeneous services. The coexistence of heterogeneous services in such a scenario is the primary focus of this study.

[2]Several studies have been conducted for D2D mode selection depending on the investigated environment and channel dynamics, however, this is out of the scope of this paper.

In particular, in our study, we focus on the coexistence of two services, i.e., the eMBB and URLLC services, and guaranteeing their technical requirements. Due to the sensitivity of URLLC services requirements, significant progress has been made in URLLC communication. In order to address their latency and reliability requirements, the packet size and the transmission time interval (TTI) are shortened, respectively [23]. Specifically for the latter, flexible waveform numerology has been enabled. This numerology enables the adjustment of sub-carrier spacings (SCS). The SCS can take a value equal to the base SCS considered in LTE, i.e., 15 KHz, or scaling it up by a factor of 2^{μ}, where $\mu = \{0, 1, 2, ...\}$ [24]. Consequently, the TTI, i.e., 1 millisecond, is shortened by 2^{μ}, satisfying the faster transmission requirement of the URLLC communication [25]. The choice of μ-numerology, consequently, the SCS and the TTI (time slot) duration are defined by 3GPP based on the application environment [25]. Moreover, the TTI can be further shortened by introducing the mini-slot timings, where each mini-slot can occupy 2, 4, or 7 symbols in the time domain. The coexistence of URLLC with other services in 5G networks occupied a significant effort from research and standardization bodies. However, it still needs to be addressed for B5G/6G networks due to the more stringent requirements: sub-millisecond latency and ultra-high reliability ($\leq 10^{-5}$ error probability) of the emergent URLLC applications [23, 26, 27].

In this paper, to enable the coexistence of eMBB and URLLC services, we leverage network slicing. Specifically, we focus on RAN slicing, to have two isolated [3] slices, each for a service type. We adopted a dynamic sharing scheme, where no pre-reservation of resources for each slice. We investigate a cooperative NOMA-based cellular system with underlay D2D communication, where each cellular user (CU)/D2D pair provides either eMBB or URLLC service, i.e., belongs to the eMBB or URLLC slice, respectively. At the level of sharing resources between CUs (i.e., using NOMA) and between CUs and D2D pairs (i.e., D2D communication underlaying cellular communication), users (CUS and D2D pairs) of different service types referring to different slices can not share the same resources. To the best of our knowledge, no existing work from the literature has addressed the joint problem

[3]Slice isolation ensures that changes to one slice do not affect the required services satisfaction of another. This can be achieved by either pre-reserving fixed resources with restricted access (fixed sharing scheme) or by defining requirements for each slice to be guaranteed without restricting resource access (dynamic sharing scheme) [28, 29].

of cellular users grouping, radio resource allocation, and D2D admission in a multi-slice CNOMA cellular system with underlay D2D communications. In particular, in this study, we first enable the grouping of CUs and D2D pairs having similar service requirements in the same slice. We then provide a three-step low-complexity approach for each slice, wherein CUs are first grouped into NOMA groups, then resource blocks are allocated to each NOMA group, and finally, the D2D pairs are admitted to share the RBs allocated to each NOMA group. We analyze the performance of the proposed solution. We use three performance metrics, namely, the overall system throughput, D2D admission ratio (i.e., the ratio of admitted D2D pairs to the maximum number of D2D pairs [4] permitted in the system), and the CUs and D2D satisfaction in terms of their technical requirements. Based on the aforementioned observations in the literature, the key contributions of this work are summarized as follows:

- We consider the coexistence of two slices, eMBB and URLLC, in a cooperative NOMA-based cellular system with underlay D2D communications. In particular, the two slices have different technical requirements: a high data rate, and a low latency and high reliability, respectively.

- We formulate an optimization problem that jointly addresses cellular users grouping, radio resource blocks allocation, and D2D admission. Its objective is to maximize the overall system throughput while satisfying the technical requirements of the two slices. We model the throughput of the eMBB CUs and D2D pairs using Shannon's evaluation (following the infinite block-length regime), while we model that of URLLC users by following the finite block-length regime (FBL) due to the considered short packet communication.

- We propose a three-step, low-complexity solution to the optimization problem: a matching theory-based approach for the cellular users grouping and the D2D admission sub-problems, and a heuristic approach for the resource blocks allocation sub-problem.

- We evaluate the system's performance in terms of the overall system throughput, the ratio of admitted D2D pairs, and the satisfaction of

[4]The maximum number of D2D pairs is pre-defined to limit the D2D-cellular interference inline with the underlay D2D literature [30, 31, 32, 33].

CUs and D2D pairs based on their corresponding slices' technical requirements.

The rest of the paper is organized as follows. Section 2 presents the state of the art. Section 3 demonstrates the system model. The problem formulation and the proposed solution are presented in Sections 4 and 5, respectively. Section 6 presents and discusses the numerical results. Finally, Section 7 concludes our work and gives some perspectives.

2. Related Works

Numerous research works have been conducted to investigate the benefits and possible application scenarios of CNOMA-enabled systems. In [34], the authors address an uplink CNOMA scenario, where a dedicated half-duplex relay connects the cellular users with the BS. The authors analyze the users' average achievable throughput and the outage probability. In [35], a downlink CNOMA system is considered with and without direct link between the BS and users, with a dedicated half-duplex relay connecting them. The authors analyze the outage probability and bit error rate under different system hardware impairments. The authors in [36] propose a CNOMA-based cognitive radio system, where a near secondary user acts as a full-duplex relay assisting a far primary user. Then, the outage probability and the average achievable throughput are analyzed. The aforementioned works [34, 35, 36] consider a three-node simple scenario: the source, the relay (near user), and the far user, excluding the realistic multi-user scenario. In this context, authors in [37, 38, 39, 40] consider a multi-users CNOMA scenario. In [37], a multi-tier CNOMA scenario is proposed, where cooperation among NOMA users is considered. Then, the outage probability and throughput are analyzed. Apart from carrying out a performance analysis only as in [34, 35, 36, 37], the authors in [38, 39, 40] optimize user pairing and power allocation to maximize the overall system throughput. A full-duplex CNOMA-based cellular system is considered in [38], and a coordinated multipoint system to mitigate the intercell interference in [39]. However, authors in [40] considered an energy-harvesting-enabled half-duplex CNOMA system. In CNOMA systems, optimizing the frequency resources allocation of CNOMA users is crucial to balance the trade-off between strong and weak users' performance, ensuring the overall performance does not deteriorate. However, this is not thoroughly investigated in the aforementioned works.

Besides, for incorporating D2D communications in cellular NOMA-based systems, different research works have been conducted as in [31, 32, 33, 41, 42, 43, 44]. In [31, 32, 41], the authors propose D2D communications underlaying cellular CNOMA-based system. In [31], the authors optimize the computing resource, power, and channel allocation to minimize the overall energy consumption and delay in a mobile edge computing system. In [32], the authors aim to maximize the overall system throughput. They optimize the admission of D2D communication undelaying a CNOMA-based cellular system, i.e., which D2D pairs share the radio resources allocated to CNOMA users while ensuring the common quality of service (QoS) requirements of both cellular users and D2D pairs. In [32], the authors found that the D2D admission rate is boosted. This is due to the improved cell-edge users' performance in the CNOMA scenario, which leads to a greater acceptance of the underlay D2D pairs without degrading the overall QoS. While in [41], the authors aim to maximize the fairness among CUs and D2D pairs alongside the overall system throughput, that's by optimizing channel and power allocations. The works conducted in [33, 42] study NOMA in the underlay D2D communications, where multiple NOMA-based D2D groups share the channel allocated to one cellular user. The authors optimize subchannel and power allocation with the aim of maximizing the overall system throughput. Differently from the aforementioned works, authors in [44, 43] study mutual SIC NOMA system between cellular users and underlay D2D pairs. In [43], only one D2D pair shares the uplink channel allocated to one uplink cellular user, while mutual SIC is applied to remove the D2D-BS and cellular users-D2D interference. They formulate a joint D2D channel and power allocation to maximize overall D2D throughput while ensuring only a common minimum throughput requirement for cellular users. In contrast to [43], in [44], authors propose a system where multiple D2D pairs share the channel allocated to two downlink cellular users. Their goal is to increase the spectral efficiency and the number of admitted D2D pairs, and keep fairness among them. To achieve this objective, they formulate an optimization problem for maximizing the minimum data rate of the D2D pairs. Therefore, they assume random cellular users pairing, and they develop a joint power and channel allocation to determine the share of channels by the D2D pairs. Their solution considers a constraint on the interference level that all cellular users would be exposed to. Meanwhile, the interference that the D2D pairs would be exposed to is neglected. Moreover, the solution misses quantifying the considered common interference constraint on the cellular users as a system-

level performance metric (i.e., data rate, latency, etc.) to satisfy. It is clear that the proposed scenarios in the aforementioned works have not addressed the coexistence of diverse service types among cellular users and D2D pairs and consequently different technical requirements and constraints.

Meanwhile, several research works studied the coexistence of different service types in NOMA-based cellular system [45, 46, 47, 48]. In [45], the authors investigated the coexistence of eMBB and URLLC services by adopting a puncturing or NOMA superposition technique for pairing an eMBB/URLLC users pair. They aim to maximize the minimum throughput of eMBB users and satisfy URLLC users' requirements by optimizing the RBs allocation of the users. In [46, 47, 48], the authors consider different RAN slices for different service types. They propose a NOMA-based mobile edge computing system, where the system users are associated into slices according to their diverse communication and computing latency requirements. The authors in [46] address the joint problem of user clustering, communication and computing resources allocation, and power control for the NOMA users. The objective of the study in [46] is to minimize the total energy consumed by the uplink NOMA users when offloading their computing tasks to the edge computing server. This minimization problem is subject to constraints imposed by the slices' latency requirements. In [47], based on the framework established in [46], a similar problem is addressed under the objective of energy-aware latency minimization for both local computing and edge offloading. However, in [48], the authors address this joint problem by offering the users hybrid access between NOMA and OMA to alleviate the induced interference of NOMA scheme. The aforementioned works [45, 46, 47, 48] have not employed cooperative communication between NOMA users to reap the benefit of CNOMA communication.

A general and clear comparison between our work and the aforementioned related works is presented in Table 1.

The next section presents the proposed system model, including the corresponding equations.

Table 1: Brief comparison with recent related works.

Ref.	**Technology**	**CUs Grouping**	**D2D Admission**	**RBs Allocation**	**eMBB-URLLC Coexist**	**URLLC throughput evaluation**	**Performance Metric**
[1]	CNOMA-Underlay D2D	✓	✓		✓	Shannon's evaluation	sum throughput; number of admitted D2D pairs
[44]	NOMA-Underlay D2D		✓	✓			D2D sum throughput; D2D mean throughput; D2D throughput fairness; D2D admission rate
[33]	NOMA-Underlay D2D		✓				sum throughput; mean throughput; outage probability
[30]	OMA-Underlay D2D	✓					sum throughput
[31]	NOMA-Underlay D2D		✓				energy consumption; total latency; total cost (weighted sum of energy consumption and latency)
[34, 35, 36, 37]	CNOMA						**[34]:**users outage probability; sum-capacity **[35]:**users and overall bit error rate; users and overall outage probability **[36]:**outage probability, sum capacity **[37]:**sum throughput; outage probability
[40]	CNOMA	✓					sum throughput

Continued on next page

Table 1 – continued from previous page

Ref.	Technology	CUs Grouping	D2D Admission	RBs Allocation	eMBB-URLLC Coexist	URLLC throughput evaluation	Performance Metric
[38]	CNOMA	✓					sum throughput; mean throughput; computation time
[39]	CNOMA	✓					sum throughput
[42]	NOMA-Underlay D2D		✓				sum throughput; number of admitted D2D pairs
[43]	NOMA-Underlay D2D		✓				D2D sum throughput; mean D2D throughput
[46, 47]	NOMA	✓		✓			[46]:energy consumption; spectral and energy efficiency; computing time fairness [47]:total latency; energy efficiency
[48]	OMA/ NOMA	✓		✓			total energy consumption
[41]	NOMA-Underlay D2D	✓	✓				sum throughput; D2D and CUs average throughput fairness
[49, 50]	OMA-Underlay D2D	✓				Shannon's evaluation	[49]:CUs, D2D and sum throughput ; number of admitted D2D pairs [50]:sum throughput; number of satisfied eMBB users
[45]	NOMA	✓		✓	✓	FBL evaluation	mean eMBB average throughput; eMBB average throughput fairness; minimum eMBB average throughput

Continued on next page

Table 1 – continued from previous page

Ref.	Technology	CUs Grouping	D2D Admission	RBs Allocation	eMBB-URLLC Coexist	URLLC throughput evaluation	Performance Metric
[51]	NOMA	✓				FBL evaluation	reliability and throughput fairness
[52]	NOMA	✓					Overall system latency (energy harvesting and transmission time)
This work	CNOMA-Underlay D2D	✓	✓	✓	✓	FBL evaluation	sum throughput; D2D admission rate; CUs and D2D satisfaction (throughput, latency and reliability)

3. System Model

In this section, the system model is described. Specifically, the network model is first explained, and the CUs and D2D pairs signal model is detailed. Then, the throughput equations of eMBB and URLLC CUs and D2D pairs are formulated following Shannon's evaluation and finite block-length evaluation, respectively. The proposed system scenario is depicted in Fig. 1, and the notations used in this paper are summarized in Table 2.

3.1. Network Model

Consider the downlink transmission scenario of a power-domain NOMA-based cellular network[5] with one BS. The set of available radio RBs is denoted by $\mathcal{R} = \{r|1 \leq r \leq |\mathcal{R}|\}$, where each RB has bandwidth B. The BS serves a set of cellular users $\mathcal{U}= \{i|1 \leq i \leq |\mathcal{U}|\}$. Denote by $\mathcal{D}= \{d|1 \leq d \leq |\mathcal{D}|\}$, the set of D2D pairs that require admission to underlay the cellular network and share the RBs allocated to the cellular users. Each D2D pair consists of a D2D transmitter and D2D receiver with a maximum transmission distance between them. The cellular users and D2D pairs are distributed in the whole cell. However, more details on their distribution can be found in Section 6. The BS, the CUs, and D2D pairs are assumed to be equipped each with

[5]Note that the inter-cell interference is assumed to be avoided using interference mitigation mechanisms.

a single antenna[6]. The CUs and D2D pairs can provide two service types: eMBB service and URLLC service. We assume that each CU and D2D pair can provide only one service type, thus the CUs and D2D pairs are divided into two classes based on their service type. To enable the coexistence of these two services, RAN slicing is adopted, and thus two isolated slices are formed as follows:

- eMBB slice: corresponds to the cellular users and D2D devices that have a minimal data rate requirement.
- URLLC slice: corresponds to the cellular users and D2D devices that have a maximal delay requirement for a reliable transmission.

For the sake of presentation, the set of the two slices is denoted by $\mathcal{S} = \{s|1 \leq s \leq |\mathcal{S}|\}$. We denote by $\mathcal{U}_s$ and $\mathcal{D}_s$, the sets of cellular users and D2D pairs associated to slice $s \in \mathcal{S}$, respectively.

Downlink NOMA is employed to provide access for CUs. Thus, multiple CUs can be allocated the same RBs to receive from the BS. These multiple CUs form a NOMA group; details about NOMA grouping of CUs will be provided later. Meanwhile, by adopting the underlay D2D communication mode, D2D devices are allowed to share the RBs allocated to each NOMA group[7]. We consider orthogonal slicing, so only CUs and D2D devices from the same slice can share the same resources. Therefore, within each slice s, the system's scenario is detailed as follows:

- The cellular users $\mathcal{U}_s$ are assumed to be grouped into different NOMA groups. The cellular users in the same NOMA group are allocated the same RBs. Due to the SIC computational complexity at CUs' receivers that grows with the number of CUs in the same NOMA group (i.e., $\mathcal{O}(.^3)$) [53], and in line with most of the NOMA literature [19, 41], we assume that each NOMA group is formed by only one weak CU and one strong CU. Specifically, the CUs are assumed to be classified into two sets: strong and weak with relatively large and small channel gains

[6]It is worth mentioning that the network model can be expanded for the scenario where the BS and the users are equipped with multiple antennas.

[7]Hereinafter, the "RBs allocated to a NOMA group" refers to the RBs allocated to the CUs that form this NOMA group.

Table 2: Notations

Notation	Definition
$\mathcal{R}$	Set of resource blocks
B	Bandwidth of each RB r
$\mathcal{S}$	Set of slices
$\mathcal{U}$	Set of CUs
$\mathcal{D}$	Set of D2D pairs
$\mathcal{U}_s$	Set of CUs in slice s
$\mathcal{U}_s^{st}$;$\mathcal{U}_s^{we}$	Set of strong; weak CUs in slice s
$\mathcal{N}_s$	Set of NOMA groups in slice s
$\mathcal{D}_s$	Set of D2D pairs in slice s
P_u^r; P_v^r	Allocated power for strong;weak user over RB r
P_c^r	Cooperative relaying power of strong user over RB r
P_d^r	D2D transmit power over RB r
R^{min}	Minimum throughput requirement of eMBB slice
τ	Target transmission time of URLLC slice
X	Packet size
L	Block-length
N_0	Noise power spectral density
q	Maximum number of D2D pairs allowed to share RBs of each NOMA group
h_u^r	Channel gain between BS and strong user u over RB r
h_v^r	Channel gain between BS and weak user v over RB r
h_d^r	Channel gain from Tx to Rx of D2D pair d over RB r
$h_{u,u}$	The channel gain of the self-interference at the strong user
$h_{*,*'}^r$	Channel gain from any transmitter $*$ to receiver $*'$ over RB r
ρ	Residual self-interference factor
ω	Imperfect SIC factor
η_i^k	Decision variable for user i to be in NOMA group k
λ_r^k	Decision variable for RB r to be allocated to NOMA group k
η_d^k	Decision variable for D2D pair d to share RBs allocated to NOMA group k
b_u, b_v, b_d	Binary satisfaction variable of strong user u; weak user v; D2D pair d
$\|\mathcal{A}\|$	Cardinality of any set $\mathcal{A}$
$\mathcal{A}\backslash\mathcal{B}$	Set difference
$\binom{a}{b}$	number of all possible combinations of b elements out of a elements, computed as $\frac{a!}{b!(a-b)!}$

to the BS, respectively[8]. Without loss of generality, the strong CUs are assumed to be close to the BS, and weak CUs are farther. We denote by $\mathcal{U}_s^{st} = \{u|1 \leq u \leq |\mathcal{U}_s^{st}|\}$ and $\mathcal{U}_s^{we} = \{v|1 \leq v \leq |\mathcal{U}_s^{we}|\}$, the sets of strong and weak cellular users, where $\mathcal{U}_s^{st} \cup \mathcal{U}_s^{we} = \mathcal{U}_s$ and $|\mathcal{U}_s^{st}| = |\mathcal{U}_s^{we}| = \frac{|\mathcal{U}_s|}{2}$. We denote by $\mathcal{N}_s = \{k|1 \leq k \leq |\mathcal{N}_s|\}$ the set of NOMA groups in slice $s \in \mathcal{S}$.

- Within each NOMA group $k \in \mathcal{N}_s$, cooperative NOMA is considered, so the strong CU acts as an in-band full-duplex relay that forwards to the weak CU its message. Due to the imperfect self-interference (SI) cancellation techniques [54], the strong CU will still be affected by a residual SI level [55], quantified by $\rho \in [0, 1]$.[9]

- For the underlay D2D communication, at maximum, q underlay D2D pairs can share the RBs allocated to each NOMA group $k \in \mathcal{N}_s$ [33, 30, 32, 31]. This is to limit the D2D-cellular interference.

Binary decision schemes are used to determine the cellular users grouping into NOMA groups, the RBs allocated to each NOMA group, and the D2D admission (i.e., which D2D pairs share the RBs allocated to each NOMA group). The binary variable η_i^k indicates whether the cellular user $i \in \mathcal{U}_s$ is associated to a NOMA group $k \in \mathcal{N}_s$, or not. Similarly, λ_r^k denotes whether RB $r \in \mathcal{R}$ is allocated to NOMA group $k \in \mathcal{N}_s$ or not, and η_d^k indicates whether D2D pair $d \in \mathcal{D}_s$ shares the RBs allocated to NOMA group $k \in \mathcal{N}_s$, or not. These variables are defined in the following equations:

$$\eta_i^k = \begin{cases} 1, \text{if } i \in \mathcal{U}_s \text{ is in NOMA group } k \\ 0, \text{otherwise} \end{cases}, \tag{1}$$

$$\lambda_r^k = \begin{cases} 1, \text{if } r \in \mathcal{R} \text{ is allocated to NOMA group } k \\ 0, \text{otherwise} \end{cases}, \tag{2}$$

[8]This classification into strong and weak sets in terms of channel gain strength is in line with the need of disparity in channel gain between users of same NOMA group for ensuring better performance and SIC success [19]. Whenever this is not the case, artificial modification of channel gains for introducing channel gain disparity is investigated, and this is out of the scope of the paper.

[9]ρ indicates the ratio of the residual SI after SI cancellation; a $\rho = 0$ indicates perfect SI cancellation, and $\rho = 1$ indicates no SI cancellation at all.

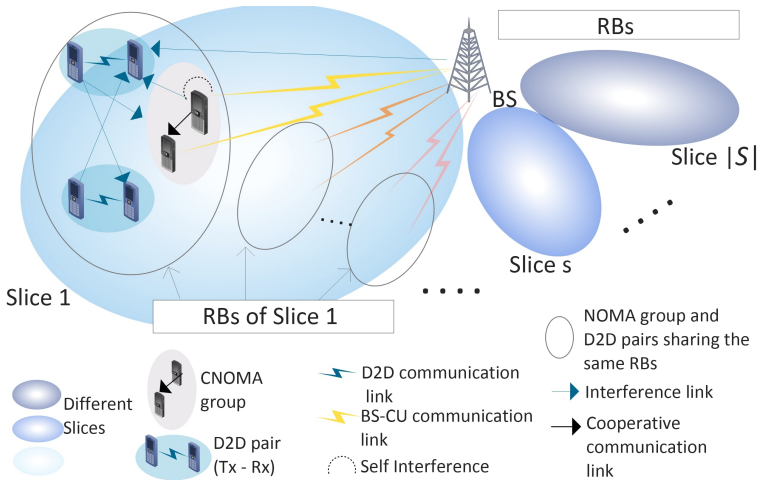


Figure 1: Multi-slices CNOMA-based cellular network with underlay D2D communications [1]

and

$$\eta_d^k = \begin{cases} 1, \text{if } d \in \mathcal{D}_s \text{ shares RBs of NOMA group } k \\ 0, \text{otherwise} \end{cases} . \tag{3}$$

3.2. Signal Model

In this section, we detail the received signals and signal-to-interference plus noise ratio (SINR) at CUs and D2D pairs.

3.2.1. CUs Signal Model

The BS superposes the signals [56] of the two CUs (i.e., a strong CU and a weak CU) and then sends the superposed signal to both users. Therefore, within a NOMA group k in slice $s \in \mathcal{S}$, the signal received by the strong cellular user, $u \in \mathcal{U}_s^{st}$, over RB r, can be given as:

$$\begin{aligned} y_u^r = & \underbrace{h_u^r(\sqrt{P_u^r}x_u + \sum\nolimits_{v \in \mathcal{U}_s^{we}} \eta_v^k \sqrt{P_v^r}x_v)}_{\text{Superposed signal from BS}} + \underbrace{n}_{\text{noise}} \\ & + \underbrace{\sqrt{\rho}h_{u,u}\sqrt{P_c^r}x_c}_{\text{Self Interference}} + \underbrace{\sum\nolimits_{d \in \mathcal{D}_s} \eta_d^k \sqrt{P_d^r}h_{d,u}^r x_d}_{\text{Interference from D2D pairs}} . \end{aligned} \tag{4}$$

where $n \sim \mathcal{CN}(0, BN_0)$ represents the additive white Gaussian noise (AWGN), and N_0 is the noise power spectral density. x_u, P_u^r, x_v, and P_v^r denote the intended messages and allocated power, over RB r, for the strong user $u \in \mathcal{U}_s^{st}$ and the weak user $v \in \mathcal{U}_s^{we}$ in the NOMA group k, respectively, such that $P_u^r \leq P_v^r$. x_c and P_c^r denote the decoded message of the weak CU at the strong CU and the relaying power to send it to the weak CU, respectively. x_d and P_d^r denote the transmit signal and the transmit power of the D2D transmitter of pair $d \in \mathcal{D}_s$, respectively.

The strong user $u \in \mathcal{U}_s^{st}$ applies SIC. First, it decodes the weak user's message x_v, following the ascending channel-based decoding order [57]. Thus, the received SINR at the strong user u to decode x_v can be given as:

$$\gamma_{u,v}^r = \frac{|h_u^r|^2 \sum_{v \in \mathcal{U}_s^{we}} \eta_v^k P_v^r}{|h_u^r|^2 P_u^r + \sum_{d \in \mathcal{D}_s} \eta_d^k |h_{d,u}^r|^2 P_d^r + \rho |h_{u,u}|^2 P_c^r + BN_0}. \tag{5}$$

After decoding and subtracting x_v, the strong user decodes its own message x_u without interference. We consider the practical SIC imperfection, represented by $\omega \in [0, 1]$. ω captures the residual interference from the weak user. Thus, the received SINR at u to decode x_u can be given as:

$$\gamma_{u,u}^r = \frac{|h_u^r|^2 P_u^r}{\sum_{d \in \mathcal{D}_s} \eta_d^k |h_{d,u}^r|^2 P_d^r + \rho |h_{u,u}|^2 P_c^r + \omega |h_u^r|^2 \sum_{v \in \mathcal{U}_s^{we}} \eta_v^k P_v^r + BN_0}. \tag{6}$$

Then, for the weak user $v \in \mathcal{U}_s^{we}$ in NOMA group k, it receives its data from both: the BS and the strong user u in the same NOMA group k. It is also affected by the interference of D2D sharing the RBs allocated to NOMA group k. Thus its received signal y_v , received SINR $\gamma_{v,v}$ to decode the data[10] transmitted by BS, and received SINR $\gamma_{v,u}$ to decode data forwarded by the strong user u can be respectively given as:

$$\begin{aligned} y_v^r = & \underbrace{h_v^r (\sqrt{P_v^r} x_v + \sum\nolimits_{u \in \mathcal{U}_s^{st}} \eta_u^k \sqrt{P_u^r} x_u)}_{\text{Superposed signal from BS}} + n \\ & + \underbrace{\sum\nolimits_{u \in \mathcal{U}_s^{st}} \eta_u^k h_{u,v}^r \sqrt{P_c^r} x_c}_{\text{Relayed signal from strong user}} + \underbrace{\sum\nolimits_{d \in \mathcal{D}_s} \eta_d^k \sqrt{P_d^r} h_{d,v}^r x_d}_{\text{Interference from D2D pairs}}, \end{aligned} \tag{7}$$

[10]Following the considered decoding order, the weak CU is the first, so it directly decodes its data considering the data of the strong CU in the superposed signal as noise.

$$\gamma_{v,v}^{r} = \frac{|h_v^r|^2 P_v^r}{|h_v^r|^2 \sum_{u \in \mathcal{U}_s^{st}} \eta_u^k P_u^r + \sum_{d \in \mathcal{D}_s} \eta_d^k |h_{d,v}^r|^2 P_d^r + BN_0}, \tag{8}$$

$$\gamma_{v,u}^{r} = \frac{\sum_{u \in \mathcal{U}_s^{st}} \eta_u^k |h_{u,v}^r|^2 P_c^r}{\sum_{d \in \mathcal{D}_s} \eta_d^k |h_{d,v}^r|^2 P_d^r + BN_0}. \tag{9}$$

The weak CU v utilizes maximal ratio combining (MRC) technique [13] to merge the signals received from both the BS and the strong user u. So its received SINR can be given as:

$$\gamma_{MRC}^{r} = \gamma_{v,u}^{r} + \gamma_{v,v}^{r}. \tag{10}$$

3.2.2. D2D Signal Model

For each D2D receiver of a D2D pair d that shares the RBs of NOMA group k, its received signal y_d^r and received SINR γ_d^r over RB r can be respectively formulated as in (11) and (12):

$$\begin{aligned} y_d^r = & \underbrace{\sqrt{P_d^r} h_d^r x_d}_{\text{D2D transmitter signal}} + \underbrace{h_{BS,d}^r \Big(\sum_{u \in \mathcal{U}_s^{st}} \eta_u^k \sqrt{P_u^r} x_u + \sum_{v \in \mathcal{U}_s^{we}} \eta_v^k \sqrt{P_v^r} x_v \Big)}_{\text{Interference from BS}} \\ & + n + \underbrace{\sum_{u \in \mathcal{U}_s^{st}} \eta_u^k h_{u,d}^r \sqrt{P_c^r} x_c}_{\text{Interference from the strong user}} + \underbrace{\sum\nolimits_{d' \in \mathcal{D}_s \setminus \{d\}} \eta_{d'}^k \sqrt{P_{d'}^r} h_{d',d}^r x_{d'}}_{\text{Interference from other D2D pairs}}, \end{aligned} \tag{11}$$

and

$$\gamma_d^r = \frac{|h_d^r|^2 P_d^r}{|h_{BS,d}^r|^2 P_k^r + \sum_{d' \in \mathcal{D}_s \setminus \{d\}} \eta_{d'}^k |h_{d',d}^r|^2 P_{d'}^r + \sum\limits_{u \in \mathcal{U}_s^{st}} \eta_u^k |h_{u,d}^r|^2 P_c^r + BN_0}, \tag{12}$$

where $P_k^r = \sum\limits_{u \in \mathcal{U}_s^{st}} \eta_u^k P_u^r + \sum\limits_{v \in \mathcal{U}_s^{we}} \eta_v^k P_v^r$, is the BS power allocated to CUs in NOMA group k over RB r. As shown in (11), the D2D receiver d is affected by interference from the BS transmitting to CUs of NOMA group k, the strong user u transmitting to the weak user v, as well as interference from the other D2D pairs sharing the RBs allocated to the same NOMA group k.

3.3. Throughput formulation of eMBB slice

For the eMBB slice, the achievable throughput of a strong user u and a weak user v in a NOMA group k can be formulated, respectively, as follows:

$$R_u^{eMBB} = B \sum\nolimits_{r \in \mathcal{R}} \lambda_r^k \log_2 \left(1 + \gamma_{u,u}^r\right), \tag{13}$$

$$R_v^{eMBB} = B \sum\nolimits_{r \in \mathcal{R}} \lambda_r^k \min(\log_2(1+\gamma_{MRC}^r), \log_2(1+\gamma_{u,v}^r)), \tag{14}$$

and the achievable throughput of a D2D pair $d \in \mathcal{D}$ sharing the RBs allocated to NOMA group k can be respectively given as follows:

$$R_d^{eMBB} = B \sum\nolimits_{r \in \mathcal{R}} \lambda_r^k log_2(1+\gamma_d^r). \tag{15}$$

The equations (13), (14) and (15) capture the maximum achievable throughput of the CUs and the D2D pairs using the Shannon capacity form [58]. However, the achievable throughput of URLLC users can not be accurately captured by the Shannon capacity form. In the next subsection, we will elaborate more and formulate the throughput of the URLLC CUs and D2D pairs.

3.4. Throughput formulation of URLLC slice

Shannon's capacity is defined as the maximum data rate such that block error probability is made low by choosing a sufficiently large packet length. Explicitly, it computes the largest transmission rate for which the communication reliability is feasible regardless of the packet length [59]. However, the hard latency and reliability requirements of the URLLC service require short-packet communication; what makes the Shannon capacity does not capture accurately the throughput of URLLC communication with the short packet length, and thus the computed latency and reliability are overestimated [23]. Therefore, we use the finite block-length regime approximation formulated by [60] to capture rate loss with respect to the Shannon rate due to URLLC short or finite-length packets. Thus the capacity of URLLC CUs and D2D pairs can be formulated as follows:

$$\begin{aligned} C^{URLLC} &= C(\gamma) - \frac{1}{ln(2)} \sqrt{\frac{V(\gamma)}{L}} Q^{-1}(\epsilon) + \mathcal{O}\left(\frac{log_2(L)}{L}\right) \\ &= log_2(1+\gamma) - \frac{1}{ln(2)} \sqrt{\frac{V(\gamma)}{\tau B \sum\limits_{r \in \mathcal{R}} \lambda_r^k}} Q^{-1}(\epsilon) + \mathcal{O}\left(\frac{log_2(\tau B \sum\limits_{r \in \mathcal{R}} \lambda_r^k)}{\tau B \sum\limits_{r \in \mathcal{R}} \lambda_r^k}\right). \end{aligned} \tag{16}$$

The first term in (16) denotes the Shannon capacity as a function of γ, where γ is computed by (6), min((5),(10)), or (12) for strong CU, weak CU,

or D2D receiver, respectively. The second term denotes the finite block-length penalty [60]. This penalty depends on the target URLLC block error probability $\epsilon = 10^{-5}$, and the block-length $L = \tau B \sum_{r \in \mathcal{R}} \lambda_r^k$, where $\tau = 0.143$ ms [11] denotes the target transmission time. $Q^{-1}(x)$ is the inverse of the Gaussian Q function, where $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{t^2}{2}} dt$, and $V = 1 - \frac{1}{(1+\gamma)^2}$ is the channel dispersion function. However the third term denotes the remainder terms in the order of $\mathcal{O}\big(\frac{log_2(L)}{L}\big)$, that can be omitted when $L \geq 50$ [60]. Thus, the achievable throughput of any URLLC CU or D2D pair can be formulated as follows:

$$R_j^{URLLC} = B \sum\nolimits_{r \in \mathcal{R}} \lambda_r^k C_j^{URLLC}, \;\; j \in \mathcal{U}; j \in \mathcal{D}. \tag{17}$$

The next section presents the formulation of our optimization problem, and the constraints on slices requirements.

4. Problem Formulation

Our problem aims to maximize the overall system throughput, satisfy the slices' technical requirements, and increase the number of admitted D2D pairs. In this section, the slices' technical requirements and the overall system throughput are defined. Then, the optimization problem is formulated.

4.1. Slices' Technical Requirements Setup

Setting the slices' technical requirements encompasses a binary decision scheme, where the metric of interest is compared with each slice's requirement, resulting in a binary output: *Satisfied (1)* or *Dissatisfied (0)*.

For the eMBB slice, the throughput of CUs and D2D pairs must exceed a predefined threshold. So, the eMBB technical requirement satisfaction can be formulated as follows :

$$R_j^{eMBB} \geq R^{min}, \quad j \in \mathcal{U}; j \in \mathcal{D} \tag{18}$$

For the URLLC slice, the CUs and D2D pairs has high reliability and low latency constraints[12]. So, the packet with X bits size must be delivered

[11] This refers for opting 0-numerology (15 kHz sub-carrier spacing), and 2-symbol transmission time interval.

[12] Note that we consider the user-plane latency, which captures the one-way transmission latency, where queuing delay and other types of delay are not considered similar to [61]. This will be considered in our future work.

within the transmission time interval, τ, reliably with ϵ error probability. Thus, the URLLC technical requirement can be formulated as follows:

$$\frac{X}{R_j^{URLLC}} \leq \tau, \quad j \in \mathcal{U}; j \in \mathcal{D}. \tag{19}$$

Therefore, the satisfaction is evaluated in the binary variable b_j as follows:

$$b_j = \begin{cases} 0, \text{ (18) or (19) does not hold} \\ 1, \text{ (18) or (19) holds} \end{cases} , j \in \mathcal{U}; j \in \mathcal{D} \tag{20}$$

4.2. Optimization Problem Formulation

The overall system throughput R_{sum} can be formulated as follows:

$$R_{sum} = \sum_{s \in \mathcal{S}} \sum_{k \in \mathcal{N}_s} \left(\sum_{u \in \mathcal{U}_s^{st}} \eta_u^k R_u + \sum_{v \in \mathcal{U}_s^{we}} \eta_v^k R_v + \sum_{d \in \mathcal{D}_s} \eta_d^k R_d \right), \tag{21}$$

where R_u, R_v and R_d are computed using (13), (14) and (15), respectively, if they are associated to the eMBB slice, or using (17), if associated to the URLLC slice.

Intuitively, our proposed problem depends on:

- Intra-NOMA group interference (inferred by η_i^k, $i \in \mathcal{U}_s$, $k \in \mathcal{N}_s$, $s \in \mathcal{S}$).
- The resource blocks allocated to each NOMA group (inferred by λ_r^k, $r \in \mathcal{R}$, $k \in \mathcal{N}_s$, $s \in \mathcal{S}$)
- Interference between the cellular users of a NOMA group and the D2D pairs sharing its RBs (inferred by η_d^k, $d \in \mathcal{D}_s$, $s \in \mathcal{S}$).

Therefore, our optimization problem can be formulated as follows:

$$
\begin{aligned}
\mathbf{P_1}: \max_{\eta,\lambda} \quad & R_{sum}(\eta,\lambda) \\
\text{s.t.} \quad & \mathbf{C}_1: \eta_i^k, \lambda_r^k, \eta_d^k \in \{0,1\}, \forall i \in \mathcal{U}_s; d \in \mathcal{D}_s; k \in \mathcal{N}_s; s \in \mathcal{S}, \\
& \mathbf{C}_2: \sum\nolimits_{k\in\mathcal{N}_s} \eta_d^k \leq 1, \quad \forall d \in \mathcal{D}_s; s \in \mathcal{S}, \\
& \mathbf{C}_3: \sum\nolimits_{k\in\mathcal{N}_s} \eta_i^k = 1, \quad \forall i \in \mathcal{U}_s; s \in \mathcal{S}, \\
& \mathbf{C}_4: \sum\nolimits_{d\in\mathcal{D}_s} \eta_d^k \leq q, \quad \forall k \in \mathcal{N}_s; s \in \mathcal{S}, \\
& \mathbf{C}_5: \sum\nolimits_{i\in\mathcal{U}_s} \eta_i^k = 2, \quad \forall k \in \mathcal{N}_s; s \in \mathcal{S}, \\
& \mathbf{C}_6: \sum_{u\in\mathcal{U}_s^{st}} \eta_u^k = 1, \sum_{v\in\mathcal{U}_s^{we}} \eta_v^k = 1, \forall k \in \mathcal{N}_s; s \in \mathcal{S}, \\
& \mathbf{C}_7: \sum\nolimits_{s\in\mathcal{S}} \sum\nolimits_{k\in\mathcal{N}_s} \sum\nolimits_{r\in\mathcal{R}} \lambda_r^k \leq |\mathcal{R}|, \\
& \mathbf{C}_8: \sum\nolimits_{s\in\mathcal{S}} \sum\nolimits_{k\in\mathcal{N}_s} \lambda_r^k = 1, \forall r \in \mathcal{R}, \\
& \mathbf{C}_9: \gamma_{u,v}^r \geq \gamma_{v,v}^r, \forall r \in \mathcal{R}, \\
& \mathbf{C}_{10}: (18), (19).
\end{aligned}
\tag{22}
$$

Constraint $\mathbf{C}_1$ ensures binary decisions of the cellular users grouping, RBs allocation and D2D admission decisions. Constraints $\mathbf{C}_2$ and $\mathbf{C}_3$ guarantee that each D2D pair d and CU i are associated to only one NOMA group k. Constraint $\mathbf{C}_4$ guarantees that at most q D2D pairs are admitted to share the RBs allocated to each NOMA group k. Constraints $\mathbf{C}_5$ and $\mathbf{C}_6$ ensure that there are only 2 CUs in each NOMA group k: one weak CU $v \in \mathcal{U}_s^{we}$, and one strong CU $u \in \mathcal{U}_s^{st}$. Constraint $\mathbf{C}_7$ guarantees that the total number of RBs allocated to NOMA groups in all slices does not exceed the total number of available RBs. Constraint $\mathbf{C}_8$ ensures that each RB r is exclusively allocated to precisely one NOMA group across the entirety of all the slices. Constraint $\mathbf{C}_9$ ensures the success of the SIC within each NOMA group k. Finally, $\mathbf{C}_{10}$ guarantees the satisfaction of the technical requirement of each slice $s \in \mathcal{S}$.

Problem $\mathbf{P_1}$ is a non-convex problem that incorporates a set of binary variables $\{\eta_i^k, \eta_d^k, \lambda_r^k\}$ that are highly coupled. Thus, $\mathbf{P_1}$ is challenging to solve and introduces expensive computational complexity to reach an optimal solution. Note that the objective function of throughput maximization in CNOMA-based systems is more complicated than that in the case of conventional NOMA systems. This is due to the achievable throughput expressions of the strong and weak users capturing the relaying and the combining

procedures. To solve $\mathbf{P_1}$, we propose to decouple it into three sequential sub-problems: *(i)* CUs grouping, *(ii)* RBs allocation, and *(iii)* D2D admission. The solutions of the three sub-problems are detailed in Section 5, where CUs grouping and D2D admission are solved using matching theory [62] and a heuristic algorithm solves the RBs allocation.

5. Proposed Solution

In this section, we present our proposed solution to the three sub-problems in three stages:

1. Grouping CUs in NOMA groups ($\eta_i^k, i \in \mathcal{U}_s, k \in \mathcal{N}_s, s \in \mathcal{S}$)
2. Allocating RBs to the NOMA groups ($\lambda_r^k, r \in \mathcal{R}, k \in \mathcal{N}_s, s \in \mathcal{S}$)
3. Admitting D2D pairs to share RBs of the NOMA groups ($\eta_d^k, d \in \mathcal{D}_s, k \in \mathcal{N}_s, s \in \mathcal{S}$),

detailed in Sections 5.1, 5.2 and 5.3, respectively.

Figure 2 presents our proposed three-stage solution, where the three stages refer to Algorithms 1, 2 and 3, respectively. Then, an analysis of the properties of the proposed solution in terms of stability, convergence, and complexity is performed in Section 5.4.

5.1. Cellular Users grouping

In the first sub-problem, $\eta_d^k, k \in \mathcal{N}_s, s \in \mathcal{S}$, is fixed to $\eta_d^{k,initial}$, i.e., initially, D2D pairs are randomly admitted to share RBs of each NOMA group k with, at maximum, q D2D pairs sharing the RBs of each NOMA group.

CUs and NOMA groups are to be matched in this sub-problem. Each CU can be matched to one NOMA group, and each NOMA group can be matched to only two CUs. Thus, a many-to-one matching, ψ_g, can map the relationship between CUs and NOMA groups. This matching can be defined as follows:

5.1.1. Many-to-One Matching Definition

Definition 1. *A many-to-one matching ψ_g is defined as a mapping of the set $\mathcal{U}_s = \mathcal{U}_s^{st} \cup \mathcal{U}_s^{we}$ to the set $\mathcal{N}_s$ if it satisfies the following conditions:*

1. $\mid \psi_g(i) \mid = 1$ *and* $\psi_g(i) \in \mathcal{N}_s$, $\forall\ i \in \mathcal{U}_s$, $s \in \mathcal{S}$;
2. $\mid \psi_g(k) \mid = 2$ *and* $\psi_g(k) \in \{\{i', i"\} | i' \in \mathcal{U}_s^{st}, i'' \in \mathcal{U}_s^{we}\}$, $\forall k \in \mathcal{N}_s$, $s \in \mathcal{S}$;

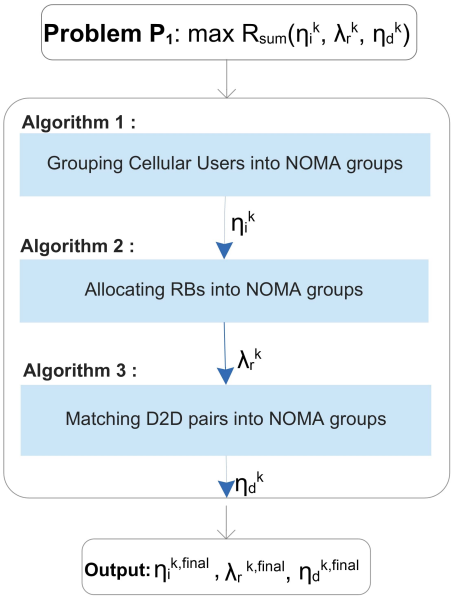


Figure 2: The proposed joint solution

3. $\psi_g(i) = k$ *if* $i \in \psi_g(k)$, $\forall\ i \in \mathcal{U}_s,\ k \in \mathcal{N}_s,\ s \in \mathcal{S}$;
4. $\psi_g(k) = i$ *if* $\psi_g(i) = k$, $\forall\ i \in \mathcal{U}_s,\ k \in \mathcal{N}_s,\ s \in \mathcal{S}$.

Condition 1) indicates that each CU $i \in \mathcal{U}_s$ is matched to only one NOMA group $k \in \mathcal{N}_s$. Condition 2) indicates that each NOMA group k can be matched to two cellular users, one strong and another weak. Conditions 3) and 4) represent the matching relationship between CU i and NOMA group k.

5.1.2. Utility Functions

Given that CUs and NOMA groups are the two players sets in this matching, each player in a set has different preferences over the players of the opposite set. These preferences are determined by the achieved utilities when matched to players from the opposite set.

To capture these preferences, we define the utility functions based on the achievable throughput and the technical requirements satisfaction. This is inline with the objective of our optimization problem and the constraints imposed by the slices' technical requirements. For any CU i, its utility function, if matched to NOMA group k under matching ψ_g, is captured in $U_i(\psi_g)$.

And for a NOMA group k, its utility function, if matched to CUs $\{i'$ and $i''\}$ under matching ψ_g, is captured in $U_k(\psi_g)$. The two utility functions are formulated as follows:

$$U_i(\psi_g) = b_i R_i, \tag{23}$$

$$U_k(\psi_g) = b_{i'} b_{i''} \left(R_{i'} + R_{i''} + \sum\nolimits_{d \in \mathcal{D}^*_{s,k}} R_d \right), \tag{24}$$

where R_i, $R_{i'}$, and $R_{i''}$ in (23) and (24) are related to the achievable throughput of the strong CU and the weak CU. $\mathcal{D}^*_{s,k}$ denotes D2D pairs that are initially randomly admitted to share RBs of NOMA group k ($\eta_d^{k,initial}$). R_d is the achievable throughput of the D2D pair d. b_i is the satisfaction variable of CU i and it is determined by (20).

5.1.3. Swapping-based Matching

Taking into consideration the interference terms in the equations of the achievable throughput, we find that the utility function of CU i on NOMA group k in (23) changes whenever a new CU is matched to the same NOMA group. So, it is not related only to the own information of CU i but also to the entire matching ψ_g. Therefore, this relationship makes the matching problem a many-to-one matching with externalities [62, 63]. Due to these externalities, CU i may swap its matched NOMA group and associate itself with another NOMA group if it finds the swap to be advantageous in terms of the utility functions. Thus, achieving a stable matching is not guaranteed under the traditional stability definitions [64]. To handle these externalities, the *pair-wise stability* definition in [65] is adopted.

First, we present the definition of *swap matching* notion in [65]. Then, we define the *pair-wise stability*.

Definition 2. *Given two CUs, i and i', matched to different NOMA groups $\psi_g(i) = k$ and $\psi_g(i') = k'$, respectively, a swap matching is defined as follows:*

$$\psi_g^{i',i} = \{\psi_g \backslash \{(i,k),(i',k')\} \cup \{(i',k),(i,k')\}, \tag{25}$$

Based on **Definition 2**, under this *swap-matching* operation, the two cellular users swap their matched NOMA groups, changing the matching from ψ_g to $\psi_g^{i',i}$. All other users and groups remain the same. Following the stability definition in [65], a *swap-matching* operation is approved if (i, i') forms a *swap-blocking pair*. The latter is defined as follows:

Definition 3. *(i, i') is a swap-blocking pair if and only if*

1. $\forall z \in \{i, i', \psi_g(i), \psi_g(i')\}$, $U_z(\psi_g^{i',i}) \geq U_z(\psi_g)$
2. $\exists z \in \{i, i', \psi_g(i), (\psi_g(i')\}$,$U_z(\psi_g^{i',i}) > U_z(\psi_g)$,

The two conditions ensure that the utilities of all agents affected by the swap operation (i.e., not only the two cellular users but also their corresponding NOMA groups) must not decrease (Condition 1) and at least one must increase (Condition 2). The latter ensures that the overall utility (i.e., the overall system throughput) is improved inline with the objective of our optimization problem. We will elaborate on this in section 5.4. Moreover, to form a *swap-blocking pair*, i, $i' \in \mathcal{U}_s^{st}$ or i, $i' \in \mathcal{U}_s^{we}; \forall s \in \mathcal{S}$. This implies that the *swap-matching* operation can occur between two strong CUs or two weak CUs. This is inline with constraints $\mathbf{C}_5$ and $\mathbf{C}_6$ in $\mathbf{P_1}$, i.e., each NOMA group is formed of one weak CU and one strong CU.

Based on the aforementioned definitions, a matching ψ_g is *pair-wise stable* if and only if there exist no *swap-blocking pairs* and no *swap-matching* operations are approved anymore. Our solution for the user grouping sub-problem using swapping-based matching theory is represented in Algorithm 1. The algorithm starts with initial matching in each slice $s \in \mathcal{S}$ by randomly pairing the CUs to form NOMA groups (lines 1-3). Then it runs the swapping-based matching (lines 4-22). In each slice, it searches for a swap-blocking pair and performs *swap-matching* operations. First, it ensures that the two users are both strong or both weak, then calculates the utility functions (lines 8-9). Then, in lines (10-15), it checks the two conditions in **Definition 3** and our problem constraints. Based on this, it approves or not, the *swap-matching* operation in **Definition 2**. The algorithm continues running until no *swap-blocking pairs* exist. It achieves a final stable matching ψ_g^{final}.

5.2. Resource blocks allocation

After having each NOMA group formed by two cellular users as an output of Algorithm 1, the second sub-problem aims to allocate the resource blocks to each NOMA group $k \in \mathcal{N}_s, s \in \mathcal{S}$.

To solve this sub-problem, $\eta_d^{k,initial}$ is randomly fixed similar to Algorithm 1. So, similarly, the initial set of admitted D2D pairs to each NOMA group $k \in \mathcal{N}_s, s \in \mathcal{S}$ is $\mathcal{D}_{s,k}^*$. The output of Algorithm 1, $\eta_i^{k,final}$ (i.e., determines the CUs in each NOMA group) is considered, then RBs are allocated to each NOMA group taking into consideration the requirements of its associated CUs.

Algorithm 1: NOMA-grouping algorithm.

Input : $\mathcal{S}$,$\mathcal{U}_s$, $\mathcal{D}^*_{s,k}$, $\psi_g = \phi$, $\forall k \in \mathcal{N}_s$, $\forall s \in \mathcal{S}$

foreach $s \in \mathcal{S}$ **do**
  As an initial matching ψ_g^0, pair CUs randomly to form NOMA groups; $\psi_g \leftarrow \psi_g^0$
end
foreach $s \in \mathcal{S}$ **do**
  repeat
    foreach $i \in \mathcal{U}_s$ **do**
      foreach $i' \in \mathcal{U}_s$ **do**
        if i, $i' \in \mathcal{U}_s^{st}$ *or* i, $i' \in \mathcal{U}_s^{we}$ **then**
          Calculate the utility functions based on (23)-(24)
          if *(i, i') satisfies the two conditions in* ***Definition 3*** *and constraints of* $\mathbf{P_1}$ **then**
            swap-matching operation is approved and i and i' swap their matches,
            $\psi_g \leftarrow \psi_g^{i',i}$
          else
            i and i' do not swap; $\psi_g \leftarrow \psi_g$
          end
        end
      end
    end
  until *No swap-blocking pairs exist*;
end

Output : $\eta_i^{k,final}, \psi_g^{final}$

The proposed solution for allocating the RBs to NOMA groups is presented in Algorithm 2. First, the algorithm calculates for each NOMA group $k \in \mathcal{N}_s, s \in \mathcal{S}$, a binary variable $\mathbf{b}_k$, equal to the product of the binary satisfaction variables (i.e., b_j in (20)) of cellular users in k and the D2D pairs sharing its RBs (line 3). The variable $\mathbf{b}_k$ is equal to: 1 if all cellular users and corresponding D2D pairs of NOMA group k are satisfied, or 0 if at least one of them is not. In the second case, a factor μ_k weighs how many elements associated to NOMA group k out of all $(\psi_g^{final}(k) \cup \mathcal{D}_{s,k}^*)$ are dissatisfied (4-6). This factor is used to prioritize the NOMA groups based on how much dissatisfied CUs and D2D pairs are associated to it. Then the algorithm runs for each RB r, first over the dissatisfied NOMA groups if there is any (line 11-18). For each NOMA group, it computes the corresponding weighted throughput based on the weighting factor μ (lines 12-15). Then, it allocates the RB r to the NOMA group k' with the maximum weighted throughput (lines 17-18).

Otherwise, when all cellular users and D2D pairs are satisfied (i.e., based on $\mathbf{b}_k$ (lines 19-27), the algorithm computes the throughput of each NOMA group when allocated RB r (lines 20-24). Then, it allocates the RB r to the NOMA group k' that achieves the maximum performance gain in the throughput after being allocated r (lines 25-26). Finally, it outputs the final RBs allocation variable $\lambda_r^{k,final}$ for each NOMA group $k \in \mathcal{N}_s, s \in \mathcal{S}$ over each RB $r \in \mathcal{R}$.

5.3. D2D Admission

The third sub-problem aims to admit the D2D pairs to share the RBs of each NOMA group. The outputs of Algorithm 1 and 2, $\eta_i^{k,final}$ and $\lambda_r^{k,final}$ respectively are taken into consideration.

Similarly to the cellular users grouping sub-problem, D2D pairs and NOMA groups are the two player sets. Each D2D pair d can be matched to only one NOMA group k, and each NOMA group k can be matched to a subset of D2D pairs $\mathcal{D}' \subset \mathcal{D}_s$ so that $|\mathcal{D}'| \leq q$. Thus a many-to-one matching, ψ_d, maps the relationship between the D2D pairs and the NOMA groups. What differs than the CUs grouping sub-problem is that some D2D pairs may be left not admitted in the system (i.e., they are not matched to any NOMA group). The utility functions of d if matched to k under matching ψ_d, and that of k if matched to $\mathcal{D}'$, under matching ψ_d, are formulated, respectively, as:

$$U_d(\psi_d) = b_d R_d \tag{26}$$

Algorithm 2: Allocation of Resource Blocks algorithm.

Input : $\mathcal{S}$, $\mathcal{D}^*_{s,k}$, ψ_g^{final}, $\eta_i^{k,final}$, $\eta_d^{k,initial}$, $\forall i \in \mathcal{U}_s, k \in \mathcal{N}_s, \forall s \in \mathcal{S}$

foreach $s \in \mathcal{S}$ **do**
  foreach $k \in \mathcal{N}_s$ **do**
    $\mathbf{b}_k = \prod_j b_j,\ j \in \psi_g^{final}(k) \cup \mathcal{D}^*_{s,k}$
    if $\mathbf{b}_k = 0$ **then**
      $\mu_k = \sum_j (1 - b_j),\ j \in \psi_g^{final}(k) \cup \mathcal{D}^*_{s,k}$
    else
      $\mu_k = 0$
    end
  end
end
foreach $r \in \mathcal{R}$ **do**
  if $\prod_{s \in \mathcal{S}} \prod_{k \in \mathcal{N}_s} \mathbf{b}_k = 0$ **then**
    foreach $s \in \mathcal{S}$ **do**
      foreach $k \in \mathcal{N}_s$ **do**
        $\mathbf{R}_k^{weighted} = \mu_k \sum_j R_j\ ,\ j \in \psi_g^{final}(k) \cup \mathcal{D}^*_{s,k}$
      end
    end
    $\lambda_r^{k'} = 1$, such that: $\mathbf{R}_{k'}^{weighted} = \max\{R_k^{weighted}, \forall k \in \mathcal{N}_s, s \in \mathcal{S}\}$
    Update $\mathbf{b}_{k'}$ and $\mu_{k'}$
  else
    foreach $s \in \mathcal{S}$ **do**
      foreach $k \in \mathcal{N}_s$ **do**
        $\mathbf{R}_k = \sum_j R_j,\ j \in \psi_g^{final}(k) \cup \mathcal{D}^*_{s,k}$
      end
    end
    $\lambda_r^{k'} = 1$, such that:
    $\mathbf{R}_{k'} - \mathbf{R'}_{k'} = \max\{\mathbf{R}_k - \mathbf{R'}_k, \forall k \in \mathcal{N}_s, s \in \mathcal{S}\}$
    $\mathbf{R'}_{k'} \leftarrow \mathbf{R}_{k'}$
  end
end

Output : $\lambda_r^{k,final}$

$$U_k(\psi_d) = b_u b_v \left(R_u + R_v + \sum\nolimits_{d' \in \mathcal{D}'} R_{d'} \right) \prod\nolimits_{d' \in \mathcal{D}'} b_d, \tag{27}$$

where b_u and b_v refer to the binary variables that indicate the satisfaction of the technical requirements of the strong and weak users, u, $v \in \psi_g^{final}(k)$, respectively. Similarly, $b_{d'}$ refers to the binary variable indicating the technical requirement satisfaction of D2D pair d'. The variables b_u, b_v and $b_{d'}$ can be computed using (20).

A similar swapping-based matching theory approach is implemented in Algorithm 3. It is based on Definitions 1, 2 and 3 but uses the aforementioned conditions of the D2D admission. Similarly, in each slice $s \in \mathcal{S}$, we aim a stable matching ψ_d^{final} while no *swap-blocking pair* (d, d') exists, $d, d' \in \mathcal{D}_s$. The outputs of Algorithm 1 and 2 are considered. Algorithm 3 starts by a random matching ψ_d^0 of the D2D pairs to NOMA groups (lines 1-3). This random matching follows a many-to-one relation and the quota q. In ψ_d^0, some D2D pairs may be initially left unmatched. Also, some NOMA groups may be initially not matched to their full quota q; let $\mathcal{N}_s^{<q} \subset \mathcal{N}_s$ denote the set of these NOMA groups. Then the algorithm starts running, and iterates until no *swap-blocking pair* exists and no *swap operation* is approved anymore (lines 8-16). Here, through a swap operation, an unmatched $d \in \mathcal{D}_s$ not only can swap with another D2D pair d', but also it can swap with an open place at a NOMA group $k \in \mathcal{N}_s^{<q}$. Finally, the algorithm outputs the final stable matching ψ_d^{final}.

5.4. Properties Analysis of the overall solution

In order to study the properties of the proposed solution (Fig. 2), its stability, convergence, and complexity are analyzed.

Theorem 1. *Convergence: Algorithms 1 and 3 converge within a finite number of iterations.*

Proof. Given the finite cardinality of the two player sets in each of the two algorithms, the number of possible swap-blocking pairs is finite, consequently a finite number of possible swap operations. Moreover following conditions of **Definition 2**, after any approved swap operation, the overall system throughput will increase. Meanwhile, the limited bandwidth (i.e., finite number of RBs) and limited power budget at the BS constrain the growth of the overall system throughput and upper-bound it in practical systems. Therefore, given that the existence of swap-blocking pairs (i.e. approval of swap

Algorithm 3: D2D admission algorithm

Input : $\mathcal{S}$, $\mathcal{D}_s$, $\eta_i^{k,final}$,$\lambda_r^{k,final}$, ψ_g^{final}, $\psi_d = \phi$, $\forall k \in \mathcal{N}_s$, $\forall s \in \mathcal{S}$

foreach $s \in \mathcal{S}$ **do**
  As an initial matching ψ_d^0, match D2D pairs randomly to NOMA groups
  $\psi_d \leftarrow \psi_d^0$
end
foreach $s \in \mathcal{S}$ **do**
  repeat
    foreach $d \in \mathcal{D}_s$ **do**
      foreach $d' \in \mathcal{D}_s \backslash \{d\} \cup \mathcal{N}_s^{<q}$ **do**
        if *(d, d') satisfies the conditions of swap-blocking pair and constraints of* $\mathbf{P_1}$ **then**
          swap-matching operation is approved and d and d' swap their matches,
          $\psi_d \leftarrow \psi_d^{d',d}$
        else
          d and d' do not swap; $\psi_d \leftarrow \psi_d$
        end
      end
    end
  until *No swap-blocking pair exists*;
end

Output : $\eta_d^{k,final}, \psi_d^{final}$

operations) is conditioned by the growth of the overall system throughput, Algorithm 1 and 3 will converge after a finite number of iterations I_u and I_d, respectively.

Theorem 2. *Stability: Algorithms 1 and 3 converge to stable matchings* ψ^{final}.

Proof. We will prove it by contradiction as follows. Based on the adopted definition of stability, the existence of swap-blocking pairs (**Definition 3**) infers that the matching is still not stable. Assume that ψ^{final} is the final matching but not stable. This means that there still exists at least one swap-

blocking pair (i, i')[13] with conditions of **Definition 3**. Consequently a swap matching operation $\psi^{i,i'}$ (**Definition 2**) will be approved, thus leading to a different matching than ψ^{final}. This contradicts the first assumption that ψ^{final} is the final matching. Therefore, this proves that the final matching is stable. If it is not stable, it is not truly the final matching, and the algorithm will keep iterating to converge to the final stable matching.

Theorem 3. *Complexity: The proposed solution has a polynomial time computational complexity.*

Proof. The "Big O Notation" estimation is used to represent the computational complexity of the algorithms. This estimation represents an asymptotic upper bound of the computational complexity [66]. The computational complexity of Algorithms 1 and 3 depends on the number of iterations I_u and I_d, respectively, that are required to achieve a stable matching, and the number of possible swap operations at each iteration. There are no closed-form expressions for I_u and I_d because it is not known how many iterations are required to ensure that no swap-blocking pairs exist anymore. However, based on **Theorem 1**, I_u and I_d are finite.

In Algorithm 1, based on **Definition 3**, each CU $i \in \mathcal{U}_s$ may form a swap-blocking pair with any other CU. Thus, at each iteration, there are $\frac{|\mathcal{U}_s|}{2} - 1$ possible swap operations for each CU $i \in \mathcal{U}_s$. Therefore, in each iteration, there are at most $|\mathcal{U}_s|\big(\frac{|\mathcal{U}_s|}{2} - 1\big)$ swap operations. The overall maximum number of swap operations is $I_u|\mathcal{U}_s|\big(\frac{|\mathcal{U}_s|}{2} - 1\big)$. So the computational complexity of Algorithm 1 is $\mathcal{O}\big(I_u|\mathcal{U}_s|^2\big)$ which is polynomial time complexity. Note that the coefficients and lower order terms are omitted as the complexity relies on the dominant terms [66].

However, the computational complexity of the exhaustive search where all possible user groupings are considered is as follows:

$$\begin{aligned}
&\mathcal{O}\left[\binom{\frac{|\mathcal{U}_s|}{2}}{1}\binom{\frac{|\mathcal{U}_s|}{2}}{1} \times \binom{\frac{|\mathcal{U}_s|}{2}-1}{1}\binom{\frac{|\mathcal{U}_s|}{2}-1}{1}\cdots\right. \\
&\left.\times\binom{\frac{|\mathcal{U}_s|}{2}-\frac{|\mathcal{U}_s|}{2}+1}{1}\binom{\frac{|\mathcal{U}_s|}{2}-\frac{|\mathcal{U}_s|}{2}+1}{1}\right] \\
&= \mathcal{O}\big(\frac{|\mathcal{U}_s|}{2}!\frac{|\mathcal{U}_s|}{2}!\big)
\end{aligned} \tag{28}$$

[13]Here (i,i') can be a swap-blocking pair in Algorithm 1 or Algorithm 3

For Algorithm 2, it allocates $|\mathcal{R}|$ RBs for $\sum_{s\in\mathcal{S}}|\mathcal{N}_s|$ NOMA groups. Thus its computational complexity is $\mathcal{O}(|\mathcal{R}||\mathcal{S}||\mathcal{N}_s|)$. However the computational complexity of exhaustive search is to find all possible ways to allocate the RBs to NOMA groups. Suppose that each NOMA group k will be allocated n_k RBs, such that $\sum_{1}^{|\mathcal{S}||\mathcal{N}_s|} n_k = |\mathcal{R}|$, and $n_k \geq 1$, $\forall k \in \mathcal{N}_s, \forall s \in \mathcal{S}$. For each possible value of n_k, the computational complexity of the exhaustive search is $\mathcal{O}\big(\frac{R!}{\prod_k n_k!}\big)$.

For Algorithm 3, the computational complexity depends on the number of iterations I_d and the number of swap operations at each iteration. For each D2D pair $d \in \mathcal{D}_s$, it may swap with another D2D pair d', or with a place of non-fully matched NOMA group. Thus at each iteration, there are at maximum $\big(|\mathcal{D}_s| + |\mathcal{N}_s^{<q}| - 1\big)$ swap operations. The overall maximum number of swap operations is $\mathrm{I}_d|\mathcal{D}_s|\big(|\mathcal{D}_s| + |\mathcal{N}_s^{<q}| - 1\big)$. So, the algorithm complexity is $\mathcal{O}\big(\mathrm{I}_d|\mathcal{D}_s\big(|\mathcal{D}_s| + |\mathcal{N}_s^{<q}|\big)\big)$

However, regarding the exhaustive search, all possible D2D-NOMA groups matching must be searched. Recall that each NOMA group can be matched to $\leq$ q D2D pairs and each D2D pair can be matched to ≤ 1 group. In this study, we consider $q = 2$ to avoid high interference levels. Since it is possible to have unmatched D2D pairs, denote by d_{adm} the number of admitted D2D pairs out of $|\mathcal{D}_s|$ i.e., to be matched to NOMA groups. Thus the computational complexity to find all possible ways to choose the D2D pairs to be admitted can be given as follows:

$$\begin{cases} \mathcal{O}\Big(\sum\limits_{d_{adm}=0}^{|\mathcal{D}_s|}\binom{|\mathcal{D}_s|}{d_{adm}}\Big) = \mathcal{O}\big(2^{|\mathcal{D}_s|}\big), \text{if} |\mathcal{D}_s| \leq 2|\mathcal{N}_s| \\ \mathcal{O}\Big(\sum\limits_{d_{adm}=0}^{2|\mathcal{N}_s|}\binom{|\mathcal{D}_s|}{d_{adm}}\Big) = \mathcal{O}\Big(\sum\limits_{d_{adm}=0}^{2|\mathcal{N}_s|}\frac{|\mathcal{D}_s|^{d_{adm}}}{d_{adm}!}\Big), \text{otherwise,} \end{cases} \tag{29}$$

since $\binom{a}{b}$ is upper bounded by $\frac{a^b}{b!}$ [67].

Then, for each of these possible ways of choosing the D2D pairs to be admitted, all possible matchings of the chosen d_{adm} D2D pairs with $|\mathcal{N}_s|$ NOMA groups have to be searched. In a particular matching, Denote by n_2 the number of NOMA groups matched to exactly 2 D2D pairs, and n_1 that of NOMA groups matched to 1 D2D pair. The rest are unmatched. Therefore the computational complexity to search all possible matchings between NOMA groups and D2D pairs is given as (29) multiplied by what follows:

$$\mathcal{O}\left[\sum_{n_2=0}\sum_{n_1=0}\binom{|\mathcal{N}_s|}{n_2}\binom{|\mathcal{N}_s|-n_2}{n_1}\prod_{i=1}^{n_2}\binom{d_{adm}-2(i-1)}{2}\right.$$
$$\left.\prod_{j=1}^{n_1}\binom{d_{adm}-2n_2-(j-1)}{1}\right] \tag{30}$$
$$=\mathcal{O}\left[\sum_{n_2=0}\sum_{n_1=0}\frac{|\mathcal{N}_s|^{(n_1+n_2)}}{n_1!n_2!}\times\frac{d_{adm}!}{2^{n_2}}\right]$$

such that $0 \leq n1+n2 \leq |\mathcal{N}_s|$, $2n_2+n_1 = d_{adm}$ and $n_2 \leq \lfloor \frac{d_{adm}}{2} \rfloor$. It is clear that the computational complexity of the exhaustive search soars since it grows exponentially as the number of D2D pairs and NOMA groups increases.

6. Numerical Results

In this section, using MATLAB, we provide the numerical results to evaluate the performance of the proposed system and solution. A uniform cellular distribution is centered around a BS with 300m coverage radius. The D2D pairs are randomly located across the cell. The strong CUs are randomly distributed within 50% of the cell radius, and the weak CUs, beyond 85% of it. Fig. 3 shows an example of the network topology with 12 CUs and 14 D2D pairs, with different colors indicating different slices. The number of NOMA groups $|\mathcal{N}_s|$ is the same $\forall s \in \mathcal{S}$, similarly $|\mathcal{D}_s|$. The channel modeling of h captures both large-scale fading, represented by path loss model PL(distance) = distance$^{-\tau}$, where $\tau = 2$ is the path loss exponent [19, 39], and small-scale fading modeled as Rayleigh fading with zero mean and unit variance [19, 39, 55]. Lastly, the SI channel coefficient is modeled as a complex random Gaussian variable with zero mean and Γ_{SI} variance [39, 55]. Unless otherwise noted, Table 3 lists the default system parameters, which are similar to those considered in the literature [19, 26, 27, 55]. All numerical results are averaged over 1000 channel realizations and cellular distributions with 95% confidence intervals.

To evaluate the performance of the proposed solution, the following metrics are considered:

- Overall system throughput that is determined according to (21).

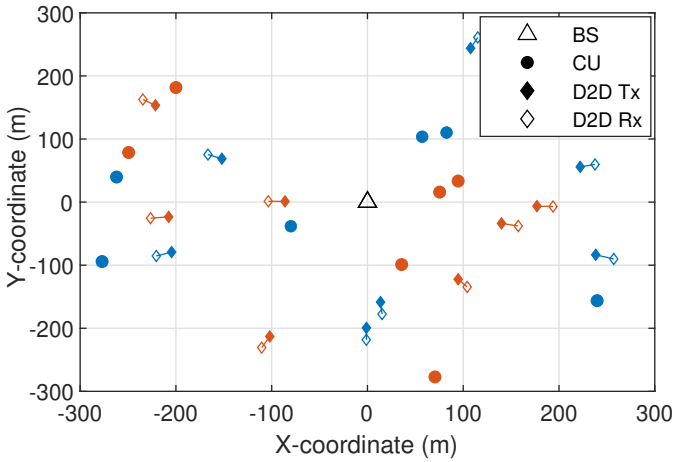


Figure 3: A snapshot of the network topology with $|\mathcal{U}| = 12$ CUs and $|\mathcal{D}| = 14$ D2D pairs

Table 3: System Parameters

Parameter	**Symbol**	**Value**
Number of NOMA groups, Number of D2D pairs; per slice	$\|\mathcal{N}_s\|$, $\|\mathcal{D}_s\|$	9, 19
Maximum power of base station	P	46 dBm
Transmit power of D2D pair	P_d	15dBm
Transmit cooperation power of strong CU	P_c	10dBm
AWGN power spectral density	N_0	-174 dBm/Hz
Sub-carrier spacing	SCS	15 KHz
Bandwidth of one RB	B	180KHz
Total Bandwidth , Number of RBs	W, $\|\mathcal{R}\|$	20 MHz, 100
eMBB minimum throughput (CUs, D2D pairs)	R^{min}	1Mbps, 0.5Mbps
URLLC target transmission time	τ	0.143 ms
URLLC packet size	X	32 bytes
URLLC block error probability	ϵ	10^{-5}
Residual SI factor	ρ	0.05
SI channel gain	Γ_{SI}	-20 dB

- CUs satisfaction ratio:
$$\frac{\sum_{s\in\mathcal{S}}\sum_{i\in\mathcal{U}_s} b_i}{|\mathcal{U}|}, \tag{31}$$
where b_i is determined according to (20).

- D2D admission ratio :
$$Z_{adm} = \frac{|\mathcal{D}_{adm}|}{Q} = \frac{\sum_{s\in\mathcal{S}}\sum_{k\in\mathcal{N}_s}\sum_{d\in\mathcal{D}_s} \eta_d^{k,final}}{\sum_{s\in\mathcal{S}} q|\mathcal{N}_s|}, \tag{32}$$
where $\mathcal{D}_{adm} \subset \mathcal{D}$ is the set of all admitted D2D pairs, and Q is the maximum number of D2D pairs allowed to be admitted.

- D2D satisfaction ratio:
$$\frac{\sum_{d\in\mathcal{D}_{adm}} b_d}{|\mathcal{D}_{adm}|}, \tag{33}$$
where b_d is determined according to (20).

6.1. Convergence of the proposed solution

First, we analyze the convergence of the proposed solution. The convergence highly depends on Algorithms 1 and 3, as the convergence of Algorithm 2 is always guaranteed. Fig. 4 shows the empirical cumulative distributive function (ECDF) of the number of approved swap operations over 1000 runs with independent channel and location realizations. The number of approved swap operations is the number of required swap operations in Algorithm 1 and 3 until reaching the final stable matchings. We compare the growth of the number of swap operations with different number of cellular users (consequently different number of NOMA groups) and number of D2D pairs. From Fig. 4, we notice that:

- Low number of swap operations is required. For example, for 9 NOMA groups and 19 D2D pairs in each slice in Fig. 4(a) and Fig. 4(b), we find that, at maximum, 36 swap operations are required for convergence.

- Lower average number of swap operations in the URLLC slice is observed. The latency and reliability requirements of the URLLC slice are less tolerant to interference, leading to less incentive by CUs and D2D pairs to swap their matches to improve the overall system throughput.

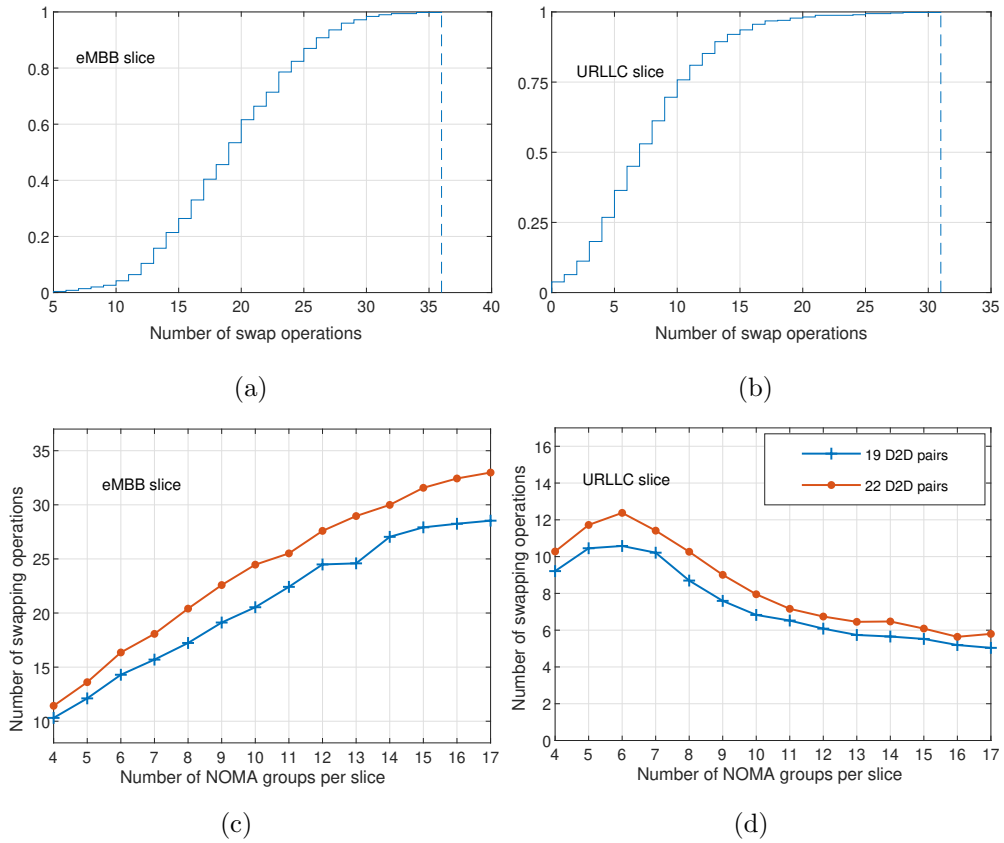
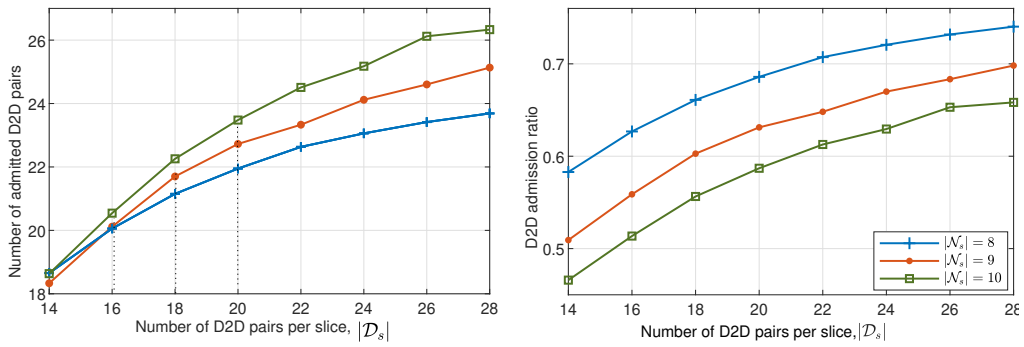


Figure 4: Convergence of the proposed solution. ECDF of the number of swap operations (a)(b); Required number of swap operations versus different number of NOMA groups and D2D pairs (c)(d).

- Comparing different network sizes (i.e. the number of NOMA groups and D2D pairs) in Fig. 4(c) and Fig. 4(d), the number of swap operations increases with the increase of the network size. This is due to the higher probability of finding more swap-blocking pairs with a higher number of NOMA groups and D2D pairs.

- The number of swap operations does not continue to increase with the number of NOMA groups. For the eMBB slice (Fig. 4(c)), it starts to stabilize at a higher number of NOMA groups, around 15, for 19 D2D pairs. For the URLLC slice (Fig. 4(d)), it starts to decrease at a higher number of NOMA groups. This is attributed to that higher number of



(a) (b)

Figure 5: The number of admitted D2D pairs (a) and the D2D admission ratio (b) with $\rho = 0.02$ under different number of D2D pairs and NOMA groups per slice.

NOMA groups within the same BS coverage induces higher interference between co-channeled CUs, and also leads to lower allocated power to each NOMA group. Given this, it is less likely to find swap-blocking pairs in the URLLC slice due to its less tolerant requirements to higher interference levels imposed by a higher number of NOMA groups.

6.2. D2D Admission under Different Network Sizes

Due to cellular-D2D interference, not all D2D pairs will be admitted to share the RBs of the CUs. Alongside the constraint on the maximum number of admitted D2D pairs over the RBs of each NOMA group (i.e., $q = 2$), the slices requirements further limit the number of admitted D2D pairs. Fig. 5 analyzes the D2D admission (number of admitted D2D pairs and the D2D admission ratio) with different numbers of NOMA groups and D2D pairs. We can notice from Fig. 5(a) that:

- The number of admitted D2D pairs $|\mathcal{D}_{adm}|$ increases with the number of available D2D pairs per slice. This is due to that the larger the set of available D2D pairs, the higher the probability of finding D2D pairs that, if admitted, do not deteriorate the individual and the overall performance.

- The rate of the increase of admitted D2D pairs starts to decline after reaching a certain value of $|\mathcal{D}_s|$. For each value of $|\mathcal{N}_s|$, observing the dotted black lines, we can see that the decline in the rate of the increase

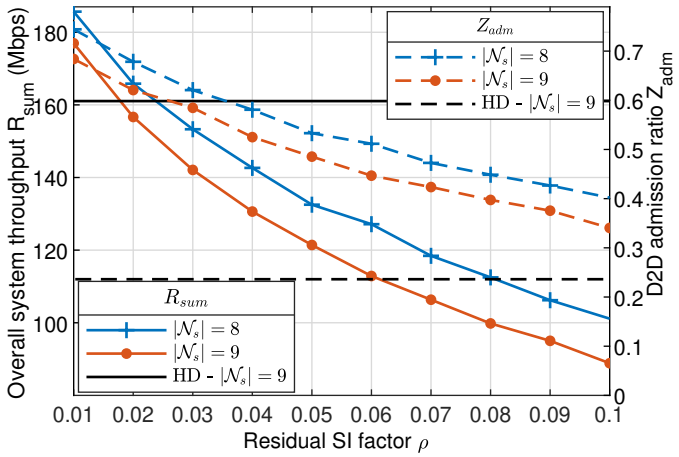


Figure 6: The overall system throughput and the D2D admission ratio under different residual SI factor values. Different numbers of NOMA groups per slice are considered.

occurs when $|\mathcal{D}_s| = q|\mathcal{N}_s|$. For instance, for $|\mathcal{N}_s| = 9$, when $|\mathcal{D}_s| \geq 18$, the number of admitted D2D pairs starts to approach a steady state.

- The number of admitted D2D pairs increases with the number of available NOMA groups per slice, $|\mathcal{N}_s|$. This is due to the underlay D2D mode where D2D pairs are admitted only by sharing RBs already allocated to available CUs. This leads to more admitted D2D pairs when more NOMA groups are available.

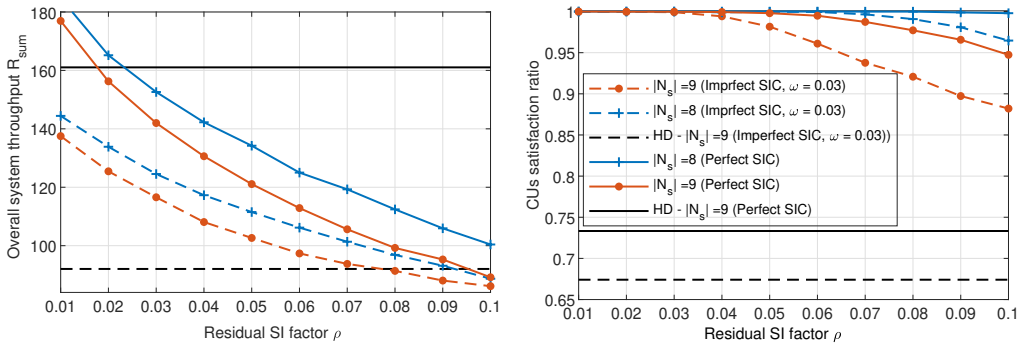
For a better representation of the D2D admission, the D2D admission ratio is utilized in Fig. 5(b) rather than the number of admitted D2D pairs. From Fig. 5(b), we can see that:

- Similarly to Fig. 5(a), the D2D admission ratio increases with the number of D2D pairs until it stabilizes as $|\mathcal{D}_s|$ approaches $q|\mathcal{N}_s|$.

- In contrary to Fig. 5(a), the D2D admission ratio decreases with $|\mathcal{N}_s|$ because higher $|\mathcal{N}_s|$ creates a denser cell under the same bandwidth and power budgets. Consequently, lowering the D2D admission ratio limit the induced interference that violates the technical requirements.

6.3. Imperfect SIC and SI

Fig. 6 and 7 analyze the performance in terms of the overall system throughput, CUs satisfaction ratio and D2D admission ratio under differ-



(a) (b)

Figure 7: The overall system throughput (a) and the CUs satisfaction ratio (b) under different residual SI factor values.

ent levels of imperfect SI cancellation within the perfect and the imperfect SIC scenarios. We compare our scenario with the half-duplex (HD) scenario, where the latter considers HD strong CUs. The results show that the higher the imperfection in the SI cancellation at the full-duplex user (i.e., higher ρ), the lower the performance and the lower the outperformance compared to the HD scenario, where the strong CUs are half-duplex relays. We can see this more pronounced in terms of R_{sum}, due to the decline of the throughputs of strong CUs impacted by the increase of SI. However, in terms of CUs satisfaction and D2D admission ratio, our scenario outperforms significantly due to the HD pre-log penalty that dissatisfies weak CUs and consequently rejects D2D admission. Moreover, comparing the perfect SIC and the imperfect SIC with $\omega = 0.03$ in Fig. 7 demonstrates that the imperfect SIC significantly deteriorates the performance. This emphasizes the significance of investigating the SIC process in NOMA-based systems to fully reap NOMA benefits.

6.4. Performance Comparison of the Proposed Solution

For the sake of comparison, the performance of our solution is compared to two other approaches from the literature, which we denote in the figures by **R-RBs** and **PJ-NOMA-1**.

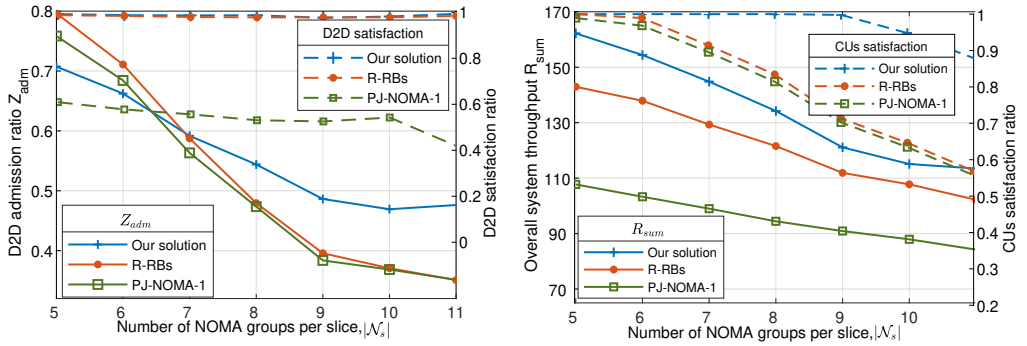
- **R-RBs** is adopted by a recent work [52]. It denotes random Eeven bandwidth division between NOMA groups. In our work, we consider the RB as the minimum bandwidth element to be allocated, so to implement even bandwidth allocation, we allocate an average of $\frac{|\mathcal{R}|}{|\mathcal{S}||\mathcal{N}_s|}$

RBs to each NOMA group . Then Algorithms 1 and 3 are implemented for CUs grouping and D2D admission, respectively. We chose **R-RBs** to compare with due to its widespread use in the literature assuming its simplicity and highest fairness [68] with the equal RBs allocation. However, our numerical results highlight the performance decline overlooked when considering it.

- **PJ-NOMA-1**, based on an approach in [44], considers fairness-based D2D RBs allocation. As such, (1) CUs are randomly grouped, (2) RBs are reserved to CUs based on their rate requirements, and then (3) each RB allocated to CUs is allocated to the D2D with the minimum achieved throughput ; this considers: (i) a maximum interference threshold on the CUs, and (ii) no constraint on D2D devices QoS. We chose **PJ-NOMA-1** to compare with because it considers fairness-based D2D admission and D2D densification, i.e., favoring higher D2D admission at the expense of performance (no D2D QoS requirement during admission). To implement **PJ-NOMA-1**, (1) CUs grouping, is achieved by Algorithm 1 in our work, (2) RBs allocation employs equal allocation due to the same requirements within each slice, and (3) D2D admission uses the D2D RBs allocation algorithm in [44]. The fairness admission condition i.e., minimum D2D throughput is replaced by the corresponding metric, whether it is an eMBB or URLLC D2D pair. The interference constraint on CUs is replaced by the corresponding slice requirement.

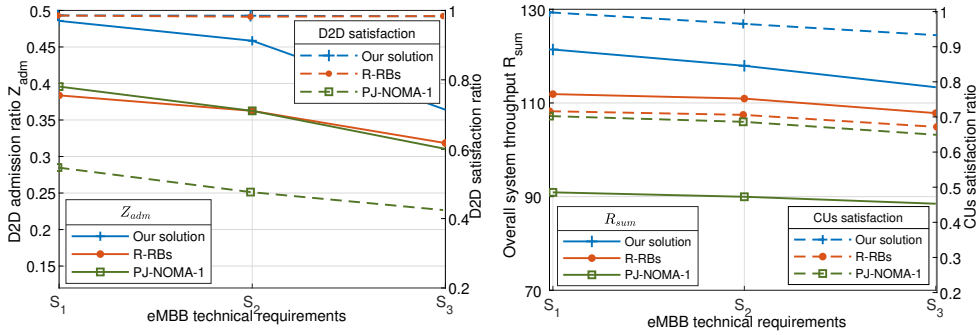
Figures 8, 9 and 10 compare our proposed solution with **PJ-NOMA-1** and **R-RBs**. In Fig. 8, we consider different network sizes by varying number of NOMA groups per slice. In Fig. 9, we consider different scenarios of eMBB technical requirements termed as S1, S2 and S3, referring to different eMBB throughput requirements of CUs and D2D pairs $(R_{CUs}^{min}, R_{d}^{min})$: (1 , 0.5), (1.25 , 0.75), and (1.5 , 1) Mbps, respectively. However, in Fig. 10, we consider different URLLC target error probabilities and packet sizes.

Moreover, to highlight the overestimation in URLLC throughput evaluation using Shannon's rate, we added another comparison baseline denoted by **Shannon's Evaluation**. We implemented our solution with URLLC Shannon's evaluation following the infinite block-length regime, as in [1, 49, 50], and we compared it with the URLLC FBL regime evaluation in our work (Fig. 10). Explicitly, in URLLC Shannon's evaluation, the throughputs of



(a) (b)

Figure 8: Performance comparison in terms of D2D admission ratio (a), D2D satisfaction ratio (a), CUs satisfaction ratio (b) and overall system throughput (b). Different numbers of NOMA groups per slice are considered.



(a) (b)

Figure 9: Performance comparison in terms of D2D admission ratio (a), D2D satisfaction ratio (a), CUs satisfaction ratio(b) and overall system throughput (b). Different minimum eMBB throughput requirements (S_1, S_2, S_3) are considered.

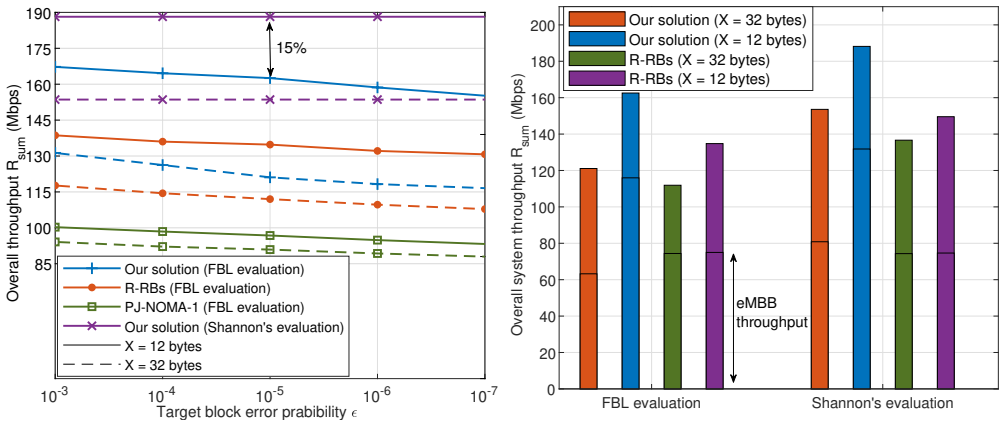


Figure 10: Performance comparison in terms of the overall system throughput (a) and the overall system throughput and eMBB sum-throughput ($\epsilon = 10^{-5}$) (b). Different URLLC target block error probability and packet sizes are considered.

URLLC CUs and D2D pairs are evaluated as the eMBB CUs and D2D pairs, using equations (13), (14) and (15), overlooking the penalty in (16).

First, in Fig. 8(a), for D2D admission ratio, at low number of NOMA groups (e.g.,$|\mathcal{N}_s| = \{5, 6\}$), **PJ-NOMA-1** and **R-RBs** achieve slightly higher than our solution. However, as $|\mathcal{N}_s|$ increases (i.e., reflecting a practical denser cellular tier), our solution outperforms them. In terms of the D2D satisfaction ratio, our proposed solution outperforms under different numbers of NOMA groups. For instance, at $|\mathcal{N}_s| = 9$, the proposed solution outperforms **R-RBs** and **PJ-NOMA-1** by 23% and 27%, respectively in terms D2D admission ratio, also it outperforms **PJ-NOMA-1** by 80 % in terms of D2D satisfaction ratio.

In Fig. 8(b), regarding the overall system throughput, our proposed solution outperforms the two other baselines for different numbers of NOMA groups. Similarly, regarding the CUs satisfaction ratio, our solution outperforms them. Moreover, the performance gap in terms of CUs satisfaction ratio is higher when the number of NOMA groups increases. This proves that our solution is more scalable even at a denser cellular tier where more interference and tightness of bandwidth and power are.

Fig. 9 shows that with more strict scenarios in terms of higher eMBB technical requirements under the same bandwidth and power resources, the D2D admission ratio (Fig.9(a)), the overall system throughput and the CUs

satisfaction ratio (Fig. 9(b)) decline. However, in the three scenarios, the performance of the proposed solution is better than that of the other comparison schemes. This makes the proposed solution more suitable even at more strict scenarios in terms of the eMBB technical requirements. Regarding the D2D satisfaction ratio, it has not decreased in our solution; however, in **PJ-NOMA-1**, it has seen a decline. This shows that our proposed solution does not prioritize the admission of more D2D pairs at the expense of their satisfaction, thus ensuring efficient resource utilization.

Finally, Fig. 10(a) shows that as the URLLC block error probability requirement tightens, R_{sum} of FBL evaluation declines. This is due to the penalty on the URLLC throughput captured in (16), necessary to maintain the required reliability level. Similarly, as the URLLC packet size X increases, R_{sum} declines due to the more stringent requirement of delivering more under the same sub-millisecond latency. However, an interesting observation from Fig. 10(b) can elaborate on this. Nonetheless, under all values of ϵ and X, our solution outperforms the other comparison schemes.

Regarding the overestimation between Shannon's evaluation and FBL regime evaluation, Fig. 10(a) shows significant gap in R_{sum}. For instance, at $\epsilon = 10^{-5}$, Shannon's evaluation overestimates R_{sum} by 15%. In the FBL regime evaluation, R_{sum} decreases, as ϵ decreases, while in Shannon's evaluation, no effect for ϵ is considered, assuming each CU and D2D pair communicates reliably at Shannon's capacity. This assumption results in increasing overestimations, from 12% to 21 %, as ϵ decreases from 10^{-3} to 10^{-7}, making this ideal assumption highly overestimating for URLLC applications with higher reliability requirements.

In Fig. 10(b), to elaborate on the decline in the overall system throughput R_{sum} as URLLC packet size X increases, we consider one value of $\epsilon = 10^{-5}$. We compare R_{sum} and the sum throughput of the eMBB slice between our proposed solution and **R-RBs**. Both show a decrease in R_{sum} as X increases. However, this decrease comes at the expense of the eMBB slice. This can be explained as follows. The increase in X demands more RBs for the URLLC slice. Unlike fixed resource sharing scheme systems, where fixed resources are pre-reserved for each slice, and no access of other slice users is allowed, RBs are dynamically shared in our system. Consequently, when X increases ($X = 32$ bytes), more RBs (than when X = 12 bytes) are allocated to the URLLC users and consequently, fewer RBs to the eMBB slice, leading to a decline in the eMBB sum throughput while still ensuring meeting the eMBB requirements. However, observing **R-RBs**, the eMBB sum throughput has

not changed due to the fixed even random RBs allocation. Consequently, URLLC users can not meet their requirements with larger packet sizes, resulting in lower CUs satisfaction and D2D admission ratios.

7. Conclusion

In this study, we have investigated the coexistence of eMBB and URLLC services in a multi-slice full-duplex CNOMA cellular system with underlay D2D communications. We formulated a joint cellular users grouping, resource blocks allocation, and D2D admission problem problem that maximizes the overall system's throughput and guarantees the slices' technical requirements. We propose a three-stage solution. In particular, we analyzed the convergence, stability and computational complexity of the proposed solution. Numerical results demonstrate its performance as a function of different SI levels and SIC imperfections, network size, and eMBB and URLLC technical requirements. Furthermore, to verify its effectiveness, we compared it with other state-of-the-art baselines and solutions. It is found to achieve higher overall system throughput, better eMBB and URLLC requirements satisfaction, and higher D2D admission, even under different strict scenarios characterized by denser networks and more stringent eMBB and URLLC requirements. As a future work, we aim to address an end-to-end latency requirement for URLLC communication, by investigating other types of delay, including queuing and computing delay alongside the transmission delay. Additionally, we aim to consider studying the impact of full-duplex mode at strong CUs on their energy consumption.

References

[1] A. Amer, S. Hoteit, J. Ben Othman, Resource allocation for enabled-network-slicing in cooperative noma-based systems with underlay d2d communications., in: 2023 IEEE International Conference on Communications (ICC), Rome, Italy, May 2023.

[2] Cisco, Cisco annual internet report (2018-2023) white paper (2020).

[3] Y. Liu, S. Zhang, X. Mu, Z. Ding, R. Schober, N. Al-Dhahir, E. Hossain, X. Shen, Evolution of noma toward next generation multiple access (ngma) for 6g, IEEE Journal on Selected Areas in Communications 40 (4) (2022) 1037–1071.

[4] S. Zhang, An overview of network slicing for 5g, IEEE Wireless Communications 26 (3) (2019) 111–117.

[5] X. Foukas, G. Patounas, A. Elmokashfi, M. Marina, Network slicing in 5g: Survey and challenges, IEEE Communications Magazine 55 (5) (2017) 94–100.

[6] Z. Wang, J. Zhang, H. Du, D. Niyato, S. Cui, B. Ai, M. Debbah, K. B. Letaief, H. V. Poor, A tutorial on extremely large-scale mimo for 6g: Fundamentals, signal processing, and applications, arXiv preprint arXiv:2307.07340 (2023).

[7] M. Chafii, L. Bariah, S. Muhaidat, M. Debbah, Twelve scientific challenges for 6g: Rethinking the foundations of communications theory, IEEE Communications Surveys & Tutorials (2023).

[8] C.-X. Wang, X. You, X. Gao, X. Zhu, Z. Li, C. Zhang, H. Wang, Y. Huang, Y. Chen, H. Haas, et al., On the road to 6g: Visions, requirements, key technologies and testbeds, IEEE Communications Surveys & Tutorials (2023).

[9] Future technology trends of terrestrial imt systems towards 2030 and beyond, Int. Telecommun. Union, Geneva, Switzerland, Rep. ITU (2022).

[10] Y. Liu, Z. Qin, M. Elkashlan, A. Nallanathan, J. McCann, Non-orthogonal multiple access in large-scale heterogeneous networks, IEEE Journal on Selected Areas in Communications 35 (12) (2017) 2667–2680.

[11] Y. Liu, Z. Ding, M. Elkashlan, J. Yuan, Nonorthogonal multiple access in large-scale underlay cognitive radio networks, IEEE Transactions on Vehicular Technology 65 (12) (2016) 10152–10157.

[12] Z. Ding, Z. Yang, P. Fan, H. V. Poor, On the performance of non-orthogonal multiple access in 5g systems with randomly deployed users, IEEE Signal Processing Letters 21 (12) (2014) 1501–1505.

[13] Y. Liu, Z. Ding, M. Elkashlan, H. V. Poor, Cooperative non-orthogonal multiple access with simultaneous wireless information and power transfer, IEEE Journal on Selected Areas in Communications 34 (4) (2016) 938–953.

[14] H. Yahya, A. Ahmed, E. Alsusa, A. Al-Dweik, Z. Ding, Error rate analysis of noma: Principles, survey and future directions, IEEE Open Journal of the Communications Society (2023).

[15] M. Vaezi, R. Schober, Z. Ding, H. V. Poor, Non-orthogonal multiple access: Common myths and critical questions, IEEE Wireless Communications 26 (5) (2019) 174–180.

[16] M. Vaezi, Z. Ding, H. V. Poor, Multiple access techniques for 5G wireless networks and beyond, Vol. 159, Springer, 2019.

[17] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, V. K. Bhargava, A survey on non-orthogonal multiple access for 5g networks: Research challenges and future trends, IEEE Journal on Selected Areas in Communications 35 (10) (2017) 2181–2195.

[18] D. Tse, P. Viswanath, Fundamentals of Wireless Communication, Cambridge University Press, 2005.

[19] Z. Ding, M. Peng, H. V. Poor, Cooperative non-orthogonal multiple access in 5g systems, IEEE Communications Letters 19 (8) (2015) 1462–1465.

[20] A. Asadi, Q. Wang, V. Mancuso, A survey on device-to-device communication in cellular networks, IEEE Communications Surveys & Tutorials 16 (4) (2014) 1801–1819.

[21] X. Dai, Z. Xiao, H. Jiang, M. Alazab, J. C. Lui, S. Dustdar, J. Liu, Task co-offloading for d2d-assisted mobile edge computing in industrial internet of things, IEEE Transactions on Industrial Informatics 19 (1) (2022) 480–490.

[22] A. Ortiz, A. Asadi, M. Engelhardt, A. Klein, M. Hollick, Cbmos: Combinatorial bandit learning for mode selection and resource allocation in d2d systems, IEEE Journal on Selected Areas in Communications 37 (10) (2019) 2225–2238.

[23] M. Bennis, M. Debbah, H. V. Poor, Ultrareliable and low-latency wireless communication: Tail, risk, and scale, Proceedings of the IEEE 106 (10) (2018) 1834–1853.

[24] A. A. Zaidi, R. Baldemair, H. Tullberg, H. Bjorkegren, L. Sundstrom, J. Medbo, C. Kilinc, I. Da Silva, Waveform and numerology to support 5g services and requirements, IEEE Communications Magazine 54 (11) (2016) 90–98.

[25] D. Maaz, A. Galindo-Serrano, S. E. Elayoubi, Urllc user plane latency performance in new radio, in: 2018 25th International Conference on Telecommunications (ICT), IEEE, 2018, pp. 225–229.

[26] M. Almekhlafi, M. A. Arfaoui, C. Assi, A. Ghrayeb, Superposition-based urllc traffic scheduling in 5g and beyond wireless networks, IEEE Transactions on Communications 70 (9) (2022) 6295–6309. doi:10.1109/TCOMM.2022.3194018.

[27] M. Almekhlafi, M. A. Arfaoui, M. Elhattab, C. Assi, A. Ghrayeb, Joint resource allocation and phase shift optimization for ris-aided embb/urllc traffic multiplexing, IEEE Transactions on Communications 70 (2) (2022) 1304–1319. doi:10.1109/TCOMM.2021.3127265.

[28] S. M. A. Kazmi, L. U. Khan, N. H. Tran, C. S. Hong, Network slicing for 5G and beyond networks, Springer, 2019.

[29] D. H. Kim, S. M. A. Kazmi, A. Ndikumana, A. Manzoor, W. Saad, C. S. Hong, Distributed radio slice allocation in wireless network virtualization: Matching theory meets auctions, IEEE Access 8 (2020) 73494–73507.

[30] Z. Ji, Z. Qin, C. G. Parini, Reconfigurable intelligent surface aided cellular networks with device-to-device users, IEEE Transactions on Communications 70 (3) (2022) 1808–1819.

[31] D. X et al., Joint computing resource, power, and channel allocations for d2d-assisted and noma-based mobile edge computing, IEEE Access 7 9243–9257, 2019.

[32] A. Amer, A. Ahmad, S. Hoteit, Resource allocation for downlink full-duplex cooperative noma-based cellular system with imperfect si cancellation and underlaying d2d communications, Sensors 21 (8) (2021) 2768.

[33] S. Moussa, A. Benslimane, R. Darazi, C. Jiang, Power allocation-based noma and underlay d2d communication for public safety users in the 5g cellular network, IEEE Systems Journal (2023).

[34] A. P. Chrysologou, N. D. Chatzidiamantis, G. K. Karagiannidis, Cooperative uplink noma in d2d communications, IEEE Communications Letters 26 (11) (2022) 2567–2571.

[35] S. Beddiaf, A. Khelil, F. Khennoufa, F. Kara, H. Kaya, X. Li, K. Rabie, H. Yanikomeroglu, A unified performance analysis of cooperative noma with practical constraints: Hardware impairment, imperfect sic and csi, IEEE Access 10 (2022) 132931–132948.

[36] S. Dhanasekaran, C. M, Performance analysis of noma in full-duplex cooperative spectrum sharing systems, IEEE Transactions on Vehicular Technology 71 (8) (2022) 9095–9100.

[37] A. S. Parihar, P. Swami, V. Bhatia, On performance of swipt enabled ppp distributed cooperative noma networks using stochastic geometry, IEEE Transactions on Vehicular Technology 71 (5) (2022) 5639–5644.

[38] P. Huu, M. A. Arfaoui, S. Sharafeddine, C. M. Assi, A. Ghrayeb, A low-complexity framework for joint user pairing and power control for cooperative noma in 5g and beyond cellular networks, IEEE Transactions on Communications 68 (11) (2020) 6737–6749.

[39] M. Elhattab, M. A. Arfaoui, C. Assi, Joint clustering and power allocation in coordinated multipoint assisted c-noma cellular networks, IEEE Transactions on Communications 70 (5) (2022) 3483–3498.

[40] M. Wu, Q. Song, L. Guo, A. Jamalipour, Joint user pairing and resource allocation in a swipt-enabled cooperative noma system, IEEE Transactions on Vehicular Technology 70 (7) (2021) 6826–6840.

[41] V. Vishnoi, I. Budhiraja, S. Gupta, N. Kumar, A deep reinforcement learning scheme for sum rate and fairness maximization among d2d pairs underlaying cellular network with noma, IEEE Transactions on Vehicular Technology (2023).

[42] J. Zhao, Y. Liu, K. K. Chai, Y. Chen, M. Elkashlan, Joint subchannel and power allocation for noma enhanced d2d communications, IEEE Transactions on Communications 65 (11) (2017) 5081–5094.

[43] A. Kilzi, J. Farah, C. A. Nour, C. Douillard, Optimal resource allocation for full-duplex iot systems underlaying cellular networks with mutual sic noma, IEEE Internet of Things Journal 8 (24) (2021) 17705–17723.

[44] L. slami, G. Mirjalily, Fairness-aware resource allocation for d2d-enabled iot in noma-based cellular networks with mutual successive interference cancellation, Physical Communication 55 (2022) 101901.

[45] Y. Prathyusha, T.-L. Sheu, Coordinated resource allocations for embb and urllc in 5g communication networks, IEEE Transactions on Vehicular Technology 71 (8) (2022) 8717–8728.

[46] M. Hossain, N. Ansari, Network slicing for noma-enabled edge computing, IEEE Transactions on Cloud Computing (2021) 1–1.

[47] M. A. Hossain, N. Ansari, Energy aware latency minimization for network slicing enabled edge computing, IEEE Transactions on Green Communications and Networking 5 (4) (2021) 2150–2159.

[48] M. A. Hossain, N. Ansari, Hybrid multiple access for network slicing aware mobile edge computing, IEEE Transactions on Cloud Computing (2023) 1–12.

[49] I. O. Sanusi, K. M. Nasr, K. Moessner, Radio resource management approaches for reliable device-to-device (d2d) communication in wireless industrial applications, IEEE transactions on cognitive communications and networking 7 (3) (2020) 905–916.

[50] A. Filali, Z. Mlika, S. Cherkaoui, A. Kobbane, Dynamic sdn-based radio access network slicing with deep reinforcement learning for urllc and embb services, IEEE Transactions on Network Science and Engineering 9 (4) (2022) 2174–2187.

[51] M. Katwe, K. Singh, C.-P. Li, Z. Ding, Ultra-high rate-reliability fairness in grant-free massive urllc noma system: Joint power and channel allocation using meta-heuristic search, IEEE Transactions on Vehicular Technology (2023).

[52] G. Li, M. Zeng, D. Mishra, L. Hao, Z. Ma, O. A. Dobre, Latency minimization for irs-aided noma mec systems with wpt-enabled iot devices, IEEE Internet of Things Journal (2023).

[53] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-Lin, Z. Wang, Non-orthogonal multiple access for 5g: solutions, challenges, opportunities, and future research trends, IEEE Communications Magazine 53 (9) (2015) 74–81.

[54] W. Shaoen, G. Hanqing, J. X., S. Z., H. W., In-band full duplex wireless communications and networking for iot devices: Progress, challenges and opportunities, Future Generation Computer Systems 92 (2019) 705–714.

[55] X. Yue, Y. Liu, S. Kang, A. Nallanathan, Z. Ding, Exploiting full/half-duplex user relaying in noma systems, IEEE Transactions on Communications 66 (2) (2018) 560–575.

[56] S. Vanka, S. Srinivasa, Z. Gong, P. Vizi, K. Stamatiou, M. Haenggi, Superposition coding strategies: Design and experimental evaluation, IEEE Transactions on Wireless Communications 11 (7) (2012) 2628–2639.

[57] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, K. Higuchi, Non-orthogonal multiple access (noma) for cellular future radio access, in: 2013 IEEE 77th vehicular technology conference (VTC Spring), IEEE, 2013, pp. 1–5.

[58] A mathematical theory of communication, The Bell system technical journal 27 (3) (1948) 379–423.

[59] G. Durisi, T. Koch, P. Popovski, Toward massive, ultrareliable, and low-latency wireless communication with short packets, Proceedings of the IEEE 104 (9) (2016) 1711–1726.

[60] Y. Polyanskiy, H. V. Poor, S. Verdú, Channel coding rate in the finite blocklength regime, IEEE Transactions on Information Theory 56 (5) (2010) 2307–2359.

[61] B. Chang, L. Li, G. Zhao, Z. Chen, M. A. Imran, Autonomous d2d transmission scheme in urllc for real-time wireless control systems, IEEE Transactions on Communications 69 (8) (2021) 5546–5558.

[62] Y. Gu, W. Saad, M. Bennis, M. Debbah, Z. Han, Matching theory for future wireless networks: fundamentals and applications, IEEE Communications Magazine 53 (5) (2015) 52–59.

[63] Z. Han, Y. Gu, W. Saad, Matching Theory for Wireless Networks, Springer, 2017.

[64] D. Gale, L. S. Shapley, College admissions and the stability of marriage, The American Mathematical Monthly 69 (1) (1962) 9–15.

[65] E. Bodine-Baron, C. Lee, A. Chong, B. Hassibi, A. Wierman, Peer effects and stability in matching markets, in: "Algorithmic Game Theory: 4th International Symposium, SAGT 2011, Amalfi, Italy, October 17-19, 2011. Proceedings 4, Springer, 2011, pp. 117–129.

[66] M. Sipser, Introduction to the Theory of Computation, Thompson Course Technology, Boston, MA, USA, 2006.

[67] T. H. Cormen, C. E. Leiserson, R. L. Rivest, C. Stein, Introduction to Algorithms. Third Edition, The MIT Press, 2009.

[68] Z. Tariq, H. Z. Khan, U. Fakhar, M. Ali, A. N. Akhtar, M. Naeem, A. Wakeel, Fairness-based user association and resource blocks allocation in satellite–terrestrial integrated networks, Physical Communication 55 (2022) 101934.