Cooperative NOMA-Based User Pairing for URLLC: A Max-Min Fairness Approach

Fateme Salehi, Naaser Neda, Mohammad-Hassan Majidi, and Hamed Ahmadi

Abstract—In this paper, cooperative non-orthogonal multiple access (C-NOMA) is considered in short packet communications with finite blocklength (FBL) codes. The performance of a decode-and-forward (DF) relaying along with selection combining (SC) and maximum ratio combining (MRC) strategies at the receiver side is examined. We explore joint user pairing and resource allocation to maximize fair throughput in a downlink (DL) scenario. In each pair, the user with a stronger channel (strong user) acts as a relay for the other one (weak user), and optimal power and blocklength are allocated to achieve max-min throughput. To this end, first, only one pair is considered, and optimal resource allocation is explored. Also, a suboptimal algorithm is suggested, which converges to a near-optimal solution. Finally, the problem is extended to a general scenario, and a suboptimal C-NOMA-based user pairing is proposed. Numerical results show that the proposed C-NOMA scheme in both SC and MRC strategies significantly improves the users' fair throughput compared to the NOMA and OMA. It is also investigated that the proposed pairing scheme based on C-NOMA outperforms the Hybrid NOMA/OMA scheme from the average throughput perspective, while the fairness index degrades slightly.

Index Terms—finite blocklength, short packet communication, URLLC, cooperative NOMA, max-min fairness, user pairing.

I. INTRODUCTION

THE ever-increasing new demands such as tactile internet, high-resolution video streaming, virtual/augmented reality, autonomous vehicles, etc., with various requirements, may be somewhat challenging in terms of reliability and latency. Unlike most of the existed mobile networks designed for traditional mobile broadband (MBB) services, Internet-of-Things (IoT) attempts to connect plentiful devices with the least human intervention. IoT applications are divided into massive machine-type communications (mMTC) and ultra-reliable low-latency communications (URLLC). The first one consists of many low-cost devices with massive connections and high battery lifetime requirements. On the other hand, URLLC requirements are most related to mission-critical services in which the importance of uninterrupted and robust data exchange is far greater than anything else.

F. Salehi is with the Faculty of Electrical and Computer Engineering, University of Birjand, Birjand, Iran (e-mail: f.salehi@birjand.ac.ir).

N. Neda *(Corresponding author)* is with the Faculty of Electrical and Computer Engineering, University of Birjand, Birjand, Iran (e-mail: nneda@birjand.ac.ir).

M.-H. Majidi is with the Faculty of Electrical and Computer Engineering, University of Birjand, Birjand, Iran (e-mail: m.majidi@birjand.ac.ir).

H. Ahmadi is with the Department of Electronic Engineering, University of York, United Kingdom (e-mail: hamed.ahmadi@york.ac.uk).

Short packets with FBL codes are considered to reduce the transmission delay and support low-latency communication. In the FBL regime communication, in contrast to Shannon's capacity for infinite blocklength, decoding error probability at the receiver is not negligible owing to short blocklength [2]. Polyanskiy et al. succeeded in deriving an exact approximation of the FBL regime's information rate at the AWGN channel [3]. Following that, research in this context developed to MIMO channel with quasi-static fading [4] and a quasi-static fading channel with retransmissions [5]. Furthermore, the effect of short packets on the spectrum sharing, and scheduling of delay-sensitive packets was considered in [6] and [7], respectively. In [8], massive MIMO adoption to maximize the achievable uplink data rate for industrial applications was advocated for both MRC and zero-forcing (ZF) receivers. In [9], the resource allocation for a secure mission-critical IoT communication system was studied under finite blocklength, and two optimization problems with the aim of weighted throughput maximization and total transmit power minimization were addressed. The authors in [10] proposed a cross-layer framework for optimizing user association, packet offloading rates, and bandwidth allocation for mission-critical IoT scenarios.

The NOMA performance in the FBL regime was studied in [11]–[14]. In [11], optimal power and blocklength allocation was considered in a high signal-to-noise ratio (SNR) scenario, and the amount of NOMA transmission delay reduction was determined compared to OMA in a closed-form. In [12], transmission rate and power allocation of the NOMA scheme were optimized to maximize the effective throughput of the strong user, while the throughput of the other user was guaranteed at a certain level. The transmitter's energy with a hybrid transmission scheme that combines the time division multiple access (TDMA) and NOMA was minimized in [13] subject to heterogeneous latency constraints at receivers. In [14], an optimal power allocation algorithm was proposed to achieve max-min throughput under energy, reliability, and delay constraints in a DL-NOMA transmission and compared with its optimal OMA counterpart.

Relaying is a well-known technique to increase capacity and reliability. In [15], relaying performance in the FBL regime was studied, and its advantages over the direct transmission were investigated. The throughput and effective capacity of a relaying system in the FBL regime were obtained in [16] at the presence of a quasi-static fading channel and some assumptions on average channel state information (CSI) at the transmitter. In [17], under the assumption of outdated CSI at the source, the authors maximized the FBL throughput of

a two-hop relaying system while guaranteeing a reliability constraint.

Ding et al. in [18] proposed the cooperative NOMA transmission scheme, a cooperative relaying technique in the NOMA system which fully exploits the prior knowledge available by applying the successive interference cancellation (SIC) strategy. Followed by that, they introduced a two-stage relay selection strategy in the C-NOMA network [19]. In [20], a buffer-aided C-NOMA scheme, where the intended users are equipped with buffers for cooperation, was proposed to adaptively select a direct or cooperative transmission mode, based on the instantaneous CSI and the buffer state. In [21], the authors proposed threshold-based selective C-NOMA, where the strong user forwards the symbols of weak user only if the signal-to-interference-plus-noise ratio (SINR) is greater than the pre-determined threshold value, to increase the data reliability of conventional C-NOMA networks. In [22], the authors investigated C-NOMA scheme in short-packet communications with flat Rayleigh fading channels and derived the average block error rate (BLER) of the central user and the cell-edge user theoretically for both SC and MRC strategies.

Optimization problems of average throughput and max-min throughput were studied in [23] with power and blocklength allocation between users under delay and consumed energy constraints by full search method with high complexity, but users' reliability was not guaranteed. In [24], Ren et al. considered optimal power and blocklength allocation in OMA, NOMA, relaying, and C-NOMA transmissions schemes to minimize the weak user's decoding error probability; meanwhile, the reliability of the strong user's performance was guaranteed at a certain level. Both [23] and [24] have considered a two-user scenario.

In [25], a joint user pairing and power allocation problem was explored in a DL-NOMA network to optimize the achievable sum rate with minimum rate constraint for each user. In [26], a two-step user-pairing scheme maximizing the achievable diversity gain for an OFDM-based relaying NOMA system with fixed-rate transmission was proposed by selecting one near user and one far user for each subcarrier where far users cannot communicate with the base station (BS) directly. Zhang et al. in [27] investigated a distance-based user pairing in the C-NOMA network, where the locations of the source and typical user are fixed, and the candidate users for pairing follow the distribution of homogeneous Poisson Point Process, and two close-to-user pairing and close-to-source pairing schemes were proposed. The authors in [28] considered user pairing policy and power control scheme jointly in a DL C-NOMA system, where the objective is maximizing the achievable sum-rate of the whole system while guaranteeing a certain quality of service (QoS) for all users. In [29], a joint user pairing and subchannel assignment algorithm was proposed in a DL C-NOMA network that pairs a strong user with a weak user and assigns them a subchannel simultaneously, while a Stackelberg game is employed to allocate power among the users by the BS. All of the works in [25]–[29] investigate the user pairing problem in conventional communication with infinite blocklength. Moreover, the last two works do not take into consideration the geometric distance between the paired nodes in the C-NOMA scheme.

In this work, we consider a DL C-NOMA network in the short packet communications scenario. It is assumed that the paired users, their channel gain difference is high. The strong user, which performs SIC and detects the weak user's data, acts as a relay. The weak user, which receives its data via BS and relay separately can implement SC or MRC to detect its data. To the best of the authors' knowledge, this is the first work to address the problem of joint user pairing, blocklength and power allocation in a critical IoT scenario.

Our main contributions in this work are summarized as follows:

1) We obtain each user's decoding error probability in the C-NOMA transmission scheme for both SC and MRC protocols with the CSI at the transmitter (CSIT) assumption. The MRC protocol is considered for the first time in the FBL regime with different blocklengths.

2) To guarantee the quality of service (QoS) of the weak user and to improve fairness, joint power and blocklength optimization is done in both NOMA and relay phases to maximize the minimum throughput of two users in different combining scenarios, under latency, reliability, and energy constraints.

3) A suboptimal solution with near-optimal performance is proposed to decrease the complexity of the optimal resource allocation, and their computational complexity is determined.

4) The problem is extended to a multi-user scenario, and a novel joint suboptimal C-NOMA-based user pairing and resource allocation scheme is proposed. Meanwhile, the simulation results show its comparable performance to the exhaustive-search optimal algorithm.

The remainder of this paper is organized as follows. In Section II, the system model and direct transmission analysis in the FBL regime are presented. Performance analysis of the C-NOMA transmission consist of SC and MRC strategies is provided in Section III. Problem formulation with a focus on one pair is considered in Section IV. The optimal and one suboptimal solution are proposed for the problem in Section V. The problem is extended to a multi-user scenario and, one user pairing scheme is proposed in Section VI. Numerical results are presented in Section VII. Finally, Section VIII concludes the paper.

II. PRELIMINARIES ISSUES

A. System Model

As shown in Fig. 1(a), the URLLC users with different QoS requirements are paired into disjoint clusters. For simplicity, we first just focus on one pair. Section VI will provide more details of user pairing. Here we consider a cooperative relaying scenario in a DL system with one BS and two NOMA users in each C-NOMA pair. In phase I, i.e., NOMA phase, BS transmits a NOMA frame of length m^{I} symbols, which consists of two users' data (N_1 bits, user 1's data and N_2 bits, user 2's data). User 1, the strong user, performs the SIC technique and decodes user 2's data and sends that to user 2 in a frame of length m^{II} symbols in phase II, i.e., relaying

phase. The instantaneous channel coefficients of BS-user 1, BS-user 2, and user 1-user 2 links representing small scale fading and large scale fading are denoted as h_1, h_2, and $h_{1,2}$, respectively. It is assumed that the channels are quasi-static Rayleigh fading. Hence, they are constant during one frame and vary independently from one frame to the next one.

According to the power domain NOMA principle, in a two-user scenario, BS transmits $\sum_{i=1}^{2} \sqrt{p_i^{\mathrm{I}}} x_i$, where x_i is the message of user i, $i \in \{1, 2\}$, and p_i^{I} refers to the allocated power of user i in phase I. So, the received signal at user i is given by $y_i^{\mathrm{I}} = (\sqrt{p_1^{\mathrm{I}}} x_1 + \sqrt{p_2^{\mathrm{I}}} x_2) h_i + n_i$, where n_i is the complex additive white Gaussian noise with variance σ^2. Without loss of generality, it is assumed that $|h_1|^2 > |h_2|^2$, and more power should be allocated to user 2. Therefore, user 1 can perform the SIC technique to remove the interference, while user 2 suffers from the interference and cannot cancel it. If x_2 is decoded correctly by user 1, it is re-encoded and transmitted (denoted by $\sqrt{p_2^{\mathrm{II}}} x_2'$). [1] Consequently, the received signal at user 2 in the relaying phase is $y_2^{\mathrm{II}} = \sqrt{p_2^{\mathrm{II}}} x_2' \, h_{1,2} + n_{1,2}$. Let p_2^{II} show the allocated power to user 2 by the relay (user 1) in phase II, and $n_{1,2}$ is the complex additive white Gaussian noise with variance σ^2. To implement this scheme, user 1 must know whether SIC is successful or not. To this end, we suppose that BS sends the channel coding information of both user 1 and user 2 to user 1 via an error-free dedicated channel. The channel coding can help to diagnose whether the decoded data is correct or not. Thus, user 1 knows whether the SIC is successful or not [24].

B. Direct Transmission Analysis in the FBL Regime

According to [3], the achievable data rate R for a finite blocklength of m symbols ($m \geq 100$), and an acceptable BLER ε, has an exact approximation as

$$R \approx C - \sqrt{\frac{V}{m}} \frac{Q^{-1}(\varepsilon)}{\ln 2} \tag{1}$$

where $C = \log_2(1 + \gamma)$ is the Shannon capacity, γ is the SNR/SINR ratio, $Q^{-1}(\cdot)$ refers to the inverse Gaussian Q-function $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^{\infty} e^{-\frac{t^2}{2}} \, dt$, and $V = 1 - (1 + \gamma)^{-2}$ is the channel dispersion. In the FBL regime, even with perfect CSI, the transmission is not error-free and the decoding error probability is given by

$$\varepsilon \approx Q(f(\gamma, R, m)). \tag{2}$$

where $f(\gamma, R, m) \triangleq \frac{(C-R)\ln 2}{\sqrt{V/m}}$.

III. PERFORMANCE ANALYSIS OF C-NOMA TRANSMISSION

It is assumed that the receivers have access to perfect CSI, and BS and each of the users have one antenna. Also, user 2 can employ various combining strategies, including SC and MRC. In phase I, user 2 directly detects x_2 by considering x_1 as interference. The decoding error probability of x_2 at user
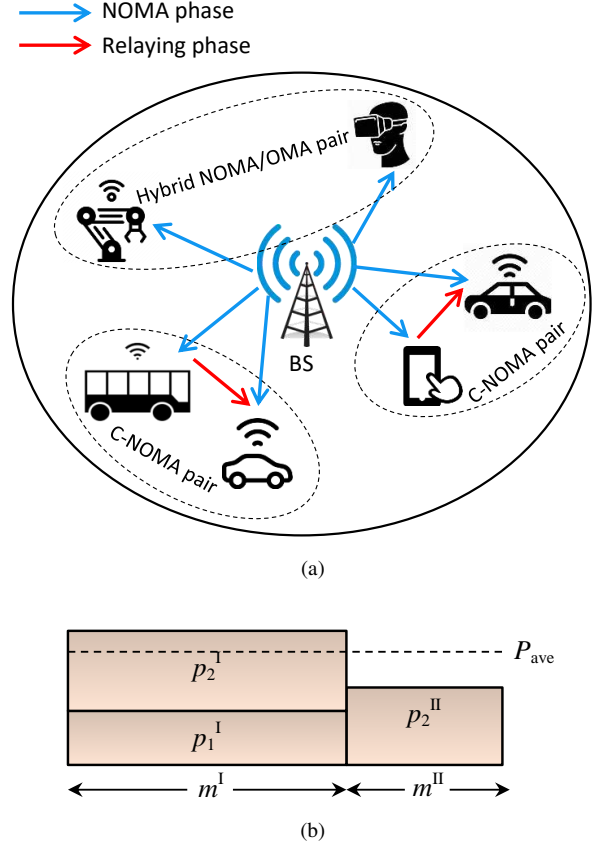

Fig. 1. (a) system model, (b) frame structure.

2 in phase I is denoted by $\varepsilon_{2,2}^{\mathrm{I}}$, which is approximated based on (2) by

$$\varepsilon_{2,2}^{\mathrm{I}} \approx Q(f(\gamma_{2,2}^{\mathrm{I}}, R_{2,2}^{\mathrm{I}}, m^{\mathrm{I}})) \tag{3}$$

where $\gamma_{2,2}^{\mathrm{I}} = p_2^{\mathrm{I}} |h_2|^2 / (p_1^{\mathrm{I}} |h_2|^2 + \phi\sigma^2)$ and $R_{2,2}^{\mathrm{I}} = N_2/m^{\mathrm{I}}$ are the received SINR and the achievable rate of user 2 related to detecting x_2 in phase I, respectively. $\phi > 1$ reflects the SNR/SINR loss due to the imperfect CSI. [2] Since x_2 is detected directly, $\varepsilon_{2,2}^{\mathrm{I}}$ is the overall error probability of user 2 in phase I, i.e., $\varepsilon_2^{\mathrm{I}} = \varepsilon_{2,2}^{\mathrm{I}}$. On the opposite, user 1 performs SIC, meaning it first decodes x_2 while treats x_1 as interference. Similarly, the decoding error probability of x_2 at user 1 in phase I, which is denoted by $\varepsilon_{1,2}^{\mathrm{I}}$, is approximated as

$$\varepsilon_{1,2}^{\mathrm{I}} \approx Q(f(\gamma_{1,2}^{\mathrm{I}}, R_{1,2}^{\mathrm{I}}, m^{\mathrm{I}})) \tag{4}$$

where $\gamma_{1,2}^{\mathrm{I}} = p_2^{\mathrm{I}} |h_1|^2 / (p_1^{\mathrm{I}} |h_1|^2 + \phi\sigma^2)$ and $R_{1,2}^{\mathrm{I}} = N_2/m^{\mathrm{I}}$ are the received SINR and the achievable rate of user 1 related to detecting x_2 in phase I, respectively. If user 1 decodes and removes x_2 successfully, then x_1 can be detected without interference. Accordingly, the decoding error probability of x_1 at user 1 in phase I, i.e., $\varepsilon_{1,1}^{\mathrm{I}}$, is denoted by

$$\varepsilon_{1,1}^{\mathrm{I}} \approx Q(f(\gamma_{1,1}^{\mathrm{I}}, R_{1,1}^{\mathrm{I}}, m^{\mathrm{I}})) \tag{5}$$

where $\gamma_{1,1}^{\mathrm{I}} = p_1^{\mathrm{I}} |h_1|^2 / \phi\sigma^2$ and $R_{1,1}^{\mathrm{I}} = N_1/m^{\mathrm{I}}$ are the received SINR and the achievable rate of user 1 related to

---

[1] One should notice that x_2 is user 2's data with rate N_2/m^{I}, while x_2' is the same data with rate N_2/m^{II}.

[2] Invoking [30], the effect of channel estimation error on data rate can be equivalent to noise enhancement, which depends on the velocity of the devices. For devices with slow or medium velocity, ϕ is close to 1.

detecting x_1 in phase I, respectively. By assuming that x_1 is detected when SIC is successful and the fact that in URLLC services, ε is usually in order of $10^{-5} \sim 10^{-9}$ [31], the overall decoding error probability at user 1 in phase I can be approximated as

$$\varepsilon_1^{\mathrm{I}} = \varepsilon_{1,2}^{\mathrm{I}} + (1 - \varepsilon_{1,2}^{\mathrm{I}})\varepsilon_{1,1}^{\mathrm{I}} \approx \varepsilon_{1,2}^{\mathrm{I}} + \varepsilon_{1,1}^{\mathrm{I}}. \qquad (6)$$

Since it is assumed that channels are half-duplex, the relayed signal is not received at user 1. Hence, the overall decoding error probability at user 1 is denoted as $\varepsilon_1 = \varepsilon_1^{\mathrm{I}}$. In contrast, the overall decoding error probability of user 2 depends on user 1 performance and thus the signal of phase II and combining strategy, where the following subsections derive the equations individually for SC and MRC strategies.

A. Selection Combining (SC)

In this protocol, user 2 does not combine the NOMA phase and relaying phase signals, but decodes transmitted messages from BS and relay (user 1) separately and selects the correctly decoded packet. First, the received message from user 1 in the relaying phase is decoded. If decoding is failed or no signal is received from user 1, then the transmitted message from BS in the NOMA phase is decoded. To differentiate the packets, the packet ID is inserted in the packet head for each device. Therefore, an error occurs when both transmissions are unsuccessful. Decoding error probability of x_2' by user 2 in phase II, i.e., $\varepsilon_{2,2}^{\mathrm{II}}$, is given by

$$\varepsilon_{2,2}^{\mathrm{II}} \approx Q(f(\gamma_{2,2}^{\mathrm{II}}, R_{2,2}^{\mathrm{II}}, m^{\mathrm{II}})) \qquad (7)$$

where $\gamma_{2,2}^{\mathrm{II}} = p_2^{\mathrm{II}}|h_{1,2}|^2/\sigma^2$ and $R_{2,2}^{\mathrm{II}} = N_2/m^{\mathrm{II}}$ are the received SNR and the achievable rate of user 2 related to detecting x_2' in phase II, respectively. One should note that the phase II signal will be transmitted if the message of user 2 is decoded correctly in phase I, so the overall decoding error probability of user 2 in phase II is approximated as

$$\varepsilon_2^{\mathrm{II}} = \varepsilon_{1,2}^{\mathrm{I}} + (1 - \varepsilon_{1,2}^{\mathrm{I}})\varepsilon_{2,2}^{\mathrm{II}} \approx \varepsilon_{1,2}^{\mathrm{I}} + \varepsilon_{2,2}^{\mathrm{II}}. \qquad (8)$$

Finally, the overall decoding error probability of user 2 in SC strategy is formulated as

$$\varepsilon_2 = \varepsilon_2^{\mathrm{I}}\varepsilon_2^{\mathrm{II}} \approx \varepsilon_{2,2}^{\mathrm{I}}(\varepsilon_{1,2}^{\mathrm{I}} + \varepsilon_{2,2}^{\mathrm{II}}). \qquad (9)$$

B. Maximum Ratio Combining (MRC)

By applying MRC protocol at user 2, since the coding rate of BS-user 2 and user 1-user 2 links are not equal, the determinative link is the bottleneck link, i.e., the link with the lowest coding rate. Therefore, the combined signal with the MRC protocol has a frame of length $m^{\mathrm{C}} = \max\{m^{\mathrm{I}}, m^{\mathrm{II}}\}$ symbols and the following SINR

$$\gamma_{2,2}^{\mathrm{C}} = \frac{m^{\mathrm{I}}}{m^{\mathrm{C}}}\gamma_{2,2}^{\mathrm{I}} + \frac{m^{\mathrm{II}}}{m^{\mathrm{C}}}\gamma_{2,2}^{\mathrm{II}}. \qquad (10)$$

The probability that user 2 fails in MRC signal decoding is given by

$$\varepsilon_{2,2}^{\mathrm{C}} \approx Q(f(\gamma_{2,2}^{\mathrm{C}}, R_{2,2}^{\mathrm{C}}, m^{\mathrm{C}})) \qquad (11)$$

where $R_{2,2}^{\mathrm{C}} = N_2/m^{\mathrm{C}}$ is the achievable rate of user 2 in the combined packet with MRC protocol.

User 2 fails when either its message is decoded correctly by none of them in phase I, or user 1 decodes x_2 correctly, but the combined signal is not decoded correctly. Hence, the overall decoding error probability of user 2 in the MRC strategy is given by

$$\varepsilon_2 = \varepsilon_{1,2}^{\mathrm{I}}\varepsilon_{2,2}^{\mathrm{I}} + (1 - \varepsilon_{1,2}^{\mathrm{I}})\varepsilon_{2,2}^{\mathrm{C}}. \qquad (12)$$

IV. PROBLEM FORMULATION

In the considered URLLC system, the two users are served with the aim of fairness during two phases with a total $D_{\max}$ symbols period. If channel feedback is available at the transmitter side, users' data rates can be set according to their instantaneous channel conditions. That being the case, a suitable criterion is max-min fairness [32]. The throughput of user i, T_i, is defined as the average bits per each channel use (or complex symbol), which is decoded correctly at the receiver;

$$T_i \triangleq \frac{m^{\mathrm{I}}}{D_{\max}}R_{i,i}^{\mathrm{I}}(1 - \varepsilon_i) \qquad (13)$$

where $1 - \varepsilon_i$ is the reliability of user i and a predefined value for each URLLC use case.

In the C-NOMA scheme, the superposition coding is performed in the NOMA phase, such that the BS enables to transmit users' signals simultaneously with different powers within a frame of length m^{I}. User 1 after decoding user 2's data, sends it in the relaying phase within a frame of length m^{II}. In Fig. 1(b) the frame structure of C-NOMA is observed. Therefore, the desired optimization problem is formulated as

$$\max_{\{p_i^{\mathrm{I}}, p_2^{\mathrm{II}}, m^j\}_{\substack{i=1,2, \\ j=\mathrm{I},\mathrm{II}}}} \min\{T_1, T_2\} \qquad (14a)$$

$$\text{s.t.} \quad m^{\mathrm{I}}\left(p_1^{\mathrm{I}} + p_2^{\mathrm{I}}\right) + m^{\mathrm{II}}p_2^{\mathrm{II}} \leq D_{\max}P_{\mathrm{ave}}, \qquad (14b)$$

$$0 < p_1^{\mathrm{I}} + p_2^{\mathrm{I}} \leq \kappa_{\mathrm{p}}P_{\mathrm{ave}}, \ p_i^{\mathrm{I}} > 0, \ i \in \{1, 2\}, \qquad (14c)$$

$$0 \leq p_2^{\mathrm{II}} \leq \kappa_{\mathrm{p}}P_{\mathrm{ave}}, \qquad (14d)$$

$$\varepsilon_i \leq \varepsilon_i^{\mathrm{th}}, \ i \in \{1, 2\}, \qquad (14e)$$

$$m^{\mathrm{I}} + m^{\mathrm{II}} = D_{\max}. \qquad (14f)$$

Optimization parameters consist of blocklength and power allocated to two users in phases I and II. Constraint (14b) indicates the system's total energy consumption budget. Constraints (14c) and (14d) are the general power constraints, where P_{ave} is the average power, and κ_{p} is the peak to average power ratio (PAPR) factor. Constraint (14e) guarantees that the decoding error probability of user i does not violate $\varepsilon_i^{\mathrm{th}}$. Moreover, the latency constraint is stated by (14f).

V. PROBLEM SOLVING

This section will solve the optimization problem in (14) for the SC and MRC strategies. To facilitate this issue, we first have to analyze the constraints and specify their optimal status. Let us first consider the constraint (14e) on the acceptable BLER of the two users. Since each URLLC use case needs specific reliability, allocating more resources to achieve a BLER lower than the required $\varepsilon_i^{\mathrm{th}}$, wastes the rare resources. Moreover, according to (1), the lower desired error probability, the lower data rate. Therefore, $\varepsilon_i = \varepsilon_i^{\mathrm{th}}$ is an optimal choice.

About constraint (14b), invoking [14, Proposition 1], the acceptable data rate (i.e., $R > 0$) in (1), is a monotonically increasing function of the corresponding SNR/SINR. Using the contradiction method, one can prove that to maximize the throughput, the energy constraint holds with equality [24], i.e., $m^{\mathrm{I}}\left(p_1^{\mathrm{I}}+p_2^{\mathrm{I}}\right)+m^{\mathrm{II}}p_2^{\mathrm{II}} = D_{\max}P_{\mathrm{ave}}$. In addition, the following proposition indicates the ratio of optimal consumed energy in two transmission phases.

Proposition 1: At the optimal solution, the total consumed energy of the two users in phase I is always greater than the consumed energy in phase II, i.e., $m^{\mathrm{I}}P_{\mathrm{sum}} > m^{\mathrm{II}}p_2^{\mathrm{II}}$, where $P_{\mathrm{sum}} \triangleq p_1^{\mathrm{I}}+p_2^{\mathrm{I}}$. (Refer to Appendix A for proof.)

Furthermore, invoking [14, Proposition 2], at the optimum point of Problem (14), throughputs of the two users are equal, i.e., $T_1 = T_2$. Following the above discussion, we provide a solution for the optimization problem in (14) with both SC and MRC strategies.

A. Optimal Design of Max-Min Fairness in C-NOMA

Since at the optimal solution $T_1 = T_2$, equation $R_{2,2}^{\mathrm{I}} = \frac{1-\varepsilon_1^{\mathrm{th}}}{1-\varepsilon_2^{\mathrm{th}}}R_{1,1}^{\mathrm{I}}$ can be derived from (13). Moreover, the message of user 2 contains the same number of bits in both phases, so it can be concluded that $R_{2,2}^{\mathrm{II}} = \frac{m^{\mathrm{I}}}{m^{\mathrm{II}}}R_{2,2}^{\mathrm{I}}$. Consequently, the optimization problem in (14) is rewritten as follows

$$\max_{\{m^{\mathrm{I}},p_1^{\mathrm{I}},p_2^{\mathrm{II}}\}} T_1 = \frac{m^{\mathrm{I}}}{D_{\max}}\left(1-\varepsilon_1^{\mathrm{th}}\right)R_{1,1}^{\mathrm{I}} \tag{15a}$$

$$\text{s.t.}\quad m^{\mathrm{I}}P_{\mathrm{sum}} + m^{\mathrm{II}}p_2^{\mathrm{II}} = D_{\max}P_{\mathrm{ave}} \tag{15b}$$

$$0 < P_{\mathrm{sum}} \le \kappa_{\mathrm{p}}P_{\mathrm{ave}},\ 0 < p_1^{\mathrm{I}} < \tfrac{P_{\mathrm{sum}}}{2} \tag{15c}$$

$$0 \le p_2^{\mathrm{II}} \le \kappa_{\mathrm{p}}P_{\mathrm{ave}},\ m^{\mathrm{I}}P_{\mathrm{sum}} > m^{\mathrm{II}}p_2^{\mathrm{II}} \tag{15d}$$

$$\varepsilon_i = \varepsilon_i^{\mathrm{th}},\ i \in \{1,2\} \tag{15e}$$

$$m^{\mathrm{I}} + m^{\mathrm{II}} = D_{\max}. \tag{15f}$$

The restriction on p_1^{I} in (15c) is applied based on the assumption that $|h_1|^2 > |h_2|^2$. So, to perform SIC correctly in the NOMA phase, it is necessary that $p_2^{\mathrm{I}} > p_1^{\mathrm{I}}$. This problem can be solved using exhaustive linear search; however, we shorten more the search range of p_1^{I} to reduce the computational complexity. The main idea can be summarized as follows:

- First, by considering user 1's decoding error probability, i.e., $\varepsilon_1 \approx \varepsilon_{1,2}^{\mathrm{I}}+\varepsilon_{1,1}^{\mathrm{I}}$, the p_1^{I} bound that guarantees $\varepsilon_1 \le \varepsilon_1^{\mathrm{th}}$ is determined. According to our previous work in [14], ε_1 is convex in p_1^{I} and at most two values hold the $\varepsilon_1(p_1^{\mathrm{I}}) = \varepsilon_1^{\mathrm{th}}$. With $R_{1,1}^{\mathrm{I}} = 0$ and constant values of m^{I} and P_{sum}, we obtain the possible solutions that keep this equality in the range of $0 < p_1^{\mathrm{I}} < \frac{P_{\mathrm{sum}}}{2}$. Clearly, $\varepsilon_{1,1}^{\mathrm{I}}$ is a monotonically decreasing function of p_1^{I}, so it is derived that $p_1^{\mathrm{I,min}} = \arg\{\varepsilon_1(p_1^{\mathrm{I}}) \approx \varepsilon_{1,1}^{\mathrm{I}}(p_1^{\mathrm{I}}) = \varepsilon_1^{\mathrm{th}}\}$. On the other hand, $\varepsilon_{1,2}^{\mathrm{I}}$ a monotonically increasing function of p_1^{I} yields to $p_1^{\mathrm{I,max}} = \arg\{\varepsilon_1(p_1^{\mathrm{I}}) \approx \varepsilon_{1,2}^{\mathrm{I}}(p_1^{\mathrm{I}}) = \varepsilon_1^{\mathrm{th}}\}$. Hence, the search region of p_1^{I} is given by $p_1^{\mathrm{I,min}} \le p_1^{\mathrm{I}} \le p_1^{\mathrm{I,max}}$.
- Since the decoding error probability is a monotonically increasing function of the transmission rate, for each value of p_1^{I} in the feasible range, $R_{1,1}^{\mathrm{I}}$ is increased until user 1's decoding error probability equals to $\varepsilon_1^{\mathrm{th}}$. One should note that $R_{1,1}^{\mathrm{I}} \le C(\gamma_{1,1}^{\mathrm{I}})$.

- Only those $p_1^{\mathrm{I,min}} \le p_1^{\mathrm{I}} \le p_1^{\mathrm{I,max}}$ that satisfy $\varepsilon_2(p_1^{\mathrm{I}}) = \varepsilon_2^{\mathrm{th}}$ could be acceptable. Since the decoding error probability of user 2 in both SC and MRC strategies, respectively in (9) and (12), are increasing function of p_1^{I}, the transmit power can be obtained using the bisection search method.
- After the full search on the values of m^{I} and P_{sum}, among the feasible solutions, the answer that maximizes T_1 is optimal.

Based on the above analysis, the algorithm for solving Problem (15) is proposed in Algorithm 1. It first determines the local maximum of T_1, i.e., $T_0^{\dagger}$, by taking constant m^{I} and checking all possible values of P_{sum} and p_1^{I}. In each iteration, the bisection search is adopted to find the desired p_1^{I}. By repeating this process on all possible m^{I} with a positive integer value, the global maximum of T_1, i.e., T_0^*, is found. Thus, using a three-dimensional (3-D) exhaustive linear search, the globally optimal solution is achieved.

Algorithm 1: Optimum Power and Blocklength Allocation Algorithm in the C-NOMA Scheme with SC/MRC Strategy

---

1 **Input**: total blocklength $D_{\max}$, overall BLER of user i $\varepsilon_i^{\mathrm{th}}$, BS average power P_{ave}, required accuracy ϵ.
2 **Output**: optimum power $p_1^{\mathrm{I}*}$, $p_2^{\mathrm{I}*}$, $p_2^{\mathrm{II}*}$, and blocklength $m^{\mathrm{I}*}$, $m^{\mathrm{II}*}$, and fair throughput $T_1 = T_2 = T_0^*$.
3 **for** $m^{\mathrm{I}} = 1 : D_{\max}$ **do**
4   **for** $P_{\mathrm{sum}} = 0 : \Delta p : \kappa_{\mathrm{p}}P_{\mathrm{ave}}$ **do**
5   Set $m^{\mathrm{II}} := D_{\max} - m^{\mathrm{I}}$ and $p_2^{\mathrm{II}} := \left(D_{\max}P_{\mathrm{ave}} - m^{\mathrm{I}}P_{\mathrm{sum}}\right)/m^{\mathrm{II}}$.
6   **if** $0 \le p_2^{\mathrm{II}} \le \kappa_{\mathrm{p}}P_{\mathrm{ave}}$ & $m^{\mathrm{I}}P_{\mathrm{sum}} \ge m^{\mathrm{II}}p_2^{\mathrm{II}}$ **then**
7   Calculate $p_1^{\mathrm{I,min}}$ and $p_1^{\mathrm{I,max}}$.
8   Set $p_1^{\mathrm{I}} := p_1^{\mathrm{I,min}}$.
9   **while** $\varepsilon_2 < \varepsilon_2^{\mathrm{th}}$ **do**
10   Set $p_1^{\mathrm{I}} := \min\left(p_1^{\mathrm{I}} + \Delta p, p_1^{\mathrm{I,max}}\right)$.
11   Find $R_{1,1}^{\mathrm{I}\dagger} = \arg\{\varepsilon_1 = \varepsilon_1^{\mathrm{th}}\}$ via bisection method with accuracy ϵ.
12   Calculate ε_2 by (9)/(12) for SC/MRC.
13   **end**
14   Set $p_1^{\mathrm{I,lb}} := p_1^{\mathrm{I}} - \Delta p$ and $p_1^{\mathrm{I,ub}} := p_1^{\mathrm{I}}$.
15   Find $p_1^{\mathrm{I}\dagger} \in \left[p_1^{\mathrm{I,lb}}, p_1^{\mathrm{I,ub}}\right]$ that satisfies $\varepsilon_2 = \varepsilon_2^{\mathrm{th}}$ via bisection method with accuracy ϵ.
16   **end**
17   **end**
18   Set $R_{1,1}^{\mathrm{I}\ddagger} := \max\left\{R_{1,1}^{\mathrm{I}\dagger} \mid \varepsilon_2 = \varepsilon_2^{\mathrm{th}}\right\}$ and $T_0^{\dagger} := \left(1-\varepsilon_1^{\mathrm{th}}\right)m^{\mathrm{I}}R_{1,1}^{\mathrm{I}\ddagger}/D_{\max}$.
19 **end**
20 Set $T_0^* := \max\{T_0^{\dagger}\}$.
21 **Return** $\{m^{\mathrm{I}*}, p_1^{\mathrm{I}*}, p_2^{\mathrm{II}*}\} = \arg\max\{T_0^{\dagger}\}$, $m^{\mathrm{II}*} = D_{\max} - m^{\mathrm{I}*}$, $p_2^{\mathrm{I}*} = \frac{\left(D_{\max}P_{\mathrm{ave}} - m^{\mathrm{II}*}p_2^{\mathrm{II}*}\right)}{m^{\mathrm{I}*}} - p_1^{\mathrm{I}*}$.

B. Suboptimal Design of Max-Min Fairness in C-NOMA

Although the search bounds of the optimum solution of Problem (15) stated in Algorithm 1 have been limited, the computational complexity is still high. Now we propose a suboptimal solution to this problem. If phase II transmission is not successful, part of the resources will go to waste, which in turn, will cause the system throughput reduction below the NOMA scheme's one. Therefore, to avoid this condition and decrease the decoding error probability in phase II, x_2' is transmitted with the maximum power, i.e., $p_2^{\mathrm{II}} = \kappa_{\mathrm{p}} P_{\mathrm{ave}}$. Hence, the summation of two users' transmit power in phase I is calculated as $P_{\mathrm{sum}} = \left[\left(D_{\max} P_{\mathrm{ave}} - m^{\mathrm{II}} p_2^{\mathrm{II}}\right)/m^{\mathrm{I}}\right]^+$, where $[x]^+ \triangleq \max\{x, 0\}$. Then, as before, the local maximum of T_1, i.e., $T_0^\dagger$, is obtained by searching on the possible values of p_1^{I} within the range of $\left[p_1^{\mathrm{I,\,min}}, p_1^{\mathrm{I,\,max}}\right]$. By repeating this process on all possible integer values of m^{I} that satisfy $m^{\mathrm{I}} P_{\mathrm{sum}} \geq m^{\mathrm{II}} p_2^{\mathrm{II}}$, the global maximum of T_1, i.e., T_0^*, is found. If $m^{\mathrm{I}} = D_{\max}$, or equivalently $m^{\mathrm{II}} = 0$, then $P_{\mathrm{sum}} = P_{\mathrm{ave}}$. In this case, signal transmission in phase II does not occur, and the C-NOMA scheme is transformed into the NOMA. This suboptimal algorithm which is a special case of Algorithm 1, needs a two-dimensional (2-D) linear search on $\{p_1^{\mathrm{I}}, m^{\mathrm{I}}\}$. The numerical results in section VII demonstrate that the performance of the suboptimal solution is slightly worse than the optimal one, while has much lower computational complexity.

C. Computational Complexity

The computational complexity of Algorithm 1 is calculated as follows. In the first step, to obtain the bounds of p_1^{I}, a linear search with complexity Ω_1 is applied. In the next step, $R_{1,1}^{\mathrm{I}}$ is derived via the bisection method with complexity around $\log_2(\varepsilon_1^{\mathrm{th}}/\epsilon)$ where ϵ is the desired accuracy. Besides, the complexity of computing ε_2 is denoted as Ω_2. This step is performed at most $K_1 = (p_1^{\mathrm{I,max}} - p_1^{\mathrm{I,min}})/\Delta p$ times where Δp is the search step, so its complexity is denoted as $K_1 \left(\log_2(\varepsilon_1^{\mathrm{th}}/\epsilon) + \Omega_2\right)$. In the last step, finding p_1^{I} via the bisection search method has complexity around $\log_2(\varepsilon_2^{\mathrm{th}}/\epsilon)$. These three steps are repeated on the possible values of P_{sum} and m^{I}, respectively $K_2 = \kappa_{\mathrm{p}} P_{\mathrm{ave}}/\Delta p$ and $D_{\max}$ times. Therefore, the worst-case complexity of Algorithm 1 is $\mathcal{O}\left(K_2 D_{\max}\left(\Omega_1 + K_1(\log_2(\varepsilon_1^{\mathrm{th}}/\epsilon) + \Omega_2) + \log_2(\varepsilon_2^{\mathrm{th}}/\epsilon)\right)\right)$.

Likewise, the computational complexity of the suboptimal algorithm is determined based on the above analysis. However, since p_2^{II} is a constant value, P_{sum} is removed from the search process. Hence, the worst-case complexity of this algorithm is $\mathcal{O}\left(D_{\max}\left(\Omega_1 + K_1(\log_2(\varepsilon_1^{\mathrm{th}}/\epsilon) + \Omega_2) + \log_2(\varepsilon_2^{\mathrm{th}}/\epsilon)\right)\right)$.

Although the number of iterations of the proposed algorithms for both SC and MRC techniques is equal, the number of basic operations related to computing the user 2's decoding error, i.e., ε_2, is different. According to (9), calculation of ε_2 in the SC technique just includes one summation and one multiplication; while, calculation of ε_2 in the MRC technique, regarding (12), requires three summations (one is due to $\gamma_{2,2}^{\mathrm{C}}$) and two multiplications.

VI. EXTENSION TO MULTI-USER SCENARIO

This section considers a more general situation shown in Fig. 1(a) when there are more than two users in a cell.

A. Problem Formulation

Let us denote the total number of users as $2K$, and the set of users as $\mathcal{K} = \{1, 2, \ldots, 2K\}$. We assume that the users' channel gains are arranged in descending order, i.e., $|h_1|^2 > |h_2|^2 > \cdots > |h_{2K}|^2$. To implement the NOMA scheme, users are grouped into some clusters. While NOMA distinguishes the users in one cluster, the various clusters become distinct by the OMA technique. Usually, in practice, to decrease the receiver's complexity, the number of users in each cluster is not considered more than four. Here we form clusters with two users and apply the C-NOMA scheme in each pair. Since for relaying, the two users need to be in the coverage area of each other; pairing is done concerning their relative locations. The number of 2-user clusters is K in the considered network, but the number of possible pairing states is completely random respecting the network topology and is denoted by Q. The throughput function of pairing in State q, where $q = 1, \ldots, Q$, is defined as follows

$$f_q\left(\mathbf{A}^q, p_i^{\mathrm{I}}, p_j^{\mathrm{II}}, m_{i,j}^{\mathrm{I}}\right) = a_{i,j}^q T_0^{i,j}; \quad i, j \in \mathcal{K}. \quad (16)$$

Let $\mathbf{A}^q = \left[a_{i,j}^q\right]_{2K \times 2K}$ be the pairing matrix in State q. Here $a_{i,j}^q$ denotes the link between users i and j in State q where

$$a_{i,j}^q = \begin{cases} 1, & \text{if users } i \text{ and } j \text{ are paired,} \\ 0, & \text{otherwise.} \end{cases} \quad (17)$$

The goal is to find the optimum pairing that maximizes the minimum throughput of the cell users. Thus, the optimization problem can be formulated as

$$\max_{q=1,\ldots,Q} \min_{\mathbf{A}^q = [a_{i,j}^q]} f_q\left(\mathbf{A}^q, p_i^{\mathrm{I}}, p_j^{\mathrm{II}}, m_{i,j}^{\mathrm{I}}\right) \quad (18a)$$

$$\text{s.t.} \quad a_{i,j}^q = a_{j,i}^q; \ i, j \in \mathcal{K} \quad (18b)$$

$$\sum_{j \in \mathcal{K} \setminus i} a_{i,j}^q \leq 1, \ i \in \mathcal{K} \quad (18c)$$

$$\sum_{i \in \mathcal{K} \setminus j} a_{i,j}^q \leq 1, \ j \in \mathcal{K}. \quad (18d)$$

Constraint (18b) shows that the pairing matrix $\mathbf{A}^q$ is symmetric. Moreover, constraints (18c) and (18d) indicate that users i and j cannot belong to more than one pair. The inter-programming problem of Problem (18) that applies the C-NOMA scheme in each pair is expressed as follows

$$T_0^{i,j} = \max_{\{m_{i,j}^{\mathrm{I}}, p_i^{\mathrm{I}}, p_j^{\mathrm{II}}\}} \min\{T_i, T_j\}, \ \forall a_{i,j}^q = 1 \quad (19a)$$

$$\text{s.t.} \quad m_{i,j}^{\mathrm{I}}\left(p_i^{\mathrm{I}} + p_j^{\mathrm{I}}\right) + m_{i,j}^{\mathrm{II}} p_j^{\mathrm{II}} = \frac{D_{\max} P_{\mathrm{ave}}}{K} \quad (19b)$$

$$0 < p_i^{\mathrm{I}} + p_j^{\mathrm{I}} \leq \kappa_{\mathrm{p}} P_{\mathrm{ave}} \quad (19c)$$

$$0 \leq p_j^{\mathrm{II}} \leq \kappa_{\mathrm{p}} P_{\mathrm{ave}} \quad (19d)$$

$$\varepsilon_i = \varepsilon_i^{\mathrm{th}}, \ \varepsilon_j = \varepsilon_j^{\mathrm{th}} \quad (19e)$$

$$m_{i,j}^{\mathrm{I}} + m_{i,j}^{\mathrm{II}} = D_{\max}. \quad (19f)$$

Here it is assumed that $|h_i|^2 > |h_j|^2$ so $p_i^{\mathrm{I}} < p_j^{\mathrm{I}}$. Constraint (19b) indicates that the total system's energy consumption is distributed equally among the pairs. For solving Problem (18),

it is needed that problem (19) is solved for all the potential pair-users in State q, i.e., $\forall a_{i,j}^q = 1$; $i, j = 1, \ldots, 2K$. Hence, to find the optimum pairing, the inter-programming problem has to be solved QK times. By an exhaustive search over all the neighboring users, every two users are paired that the minimum achieved throughput in the cell is maximized. The complexity of the exhaustive search (i.e., the number of iterations needed to find the optimal pairing) is almost high, resulting in excessive scheduling delay with a large number of users. To alleviate the computational complexity, a suboptimal pairing algorithm is proposed in the following subsection.

B. The proposed C-NOMA pairing

Here, a suboptimal solution for Problem (18) is proposed. The objective is to maximize the throughput of the weakest user among all $2K$ users by allocating them into different pairs according to the geographic locations. To implement the proposed user pairing, the graph matrix of the network topology has to be obtained first. For this purpose, each user ought to find all the users in its coverage area with radius r_0. Since the aim is leveraging C-NOMA to increase reliability and system capacity, the priority is with C-NOMA pairs starting from the weakest user. Users that are far from others and do not have a chance to exploit the C-NOMA technique use NOMA or OMA instead, depending on their channel condition. Finally, the users that have not been scheduled in C-NOMA pairs are rearranged to form hybrid NOMA/OMA pairs, as will be described in the following subsection. Algorithm 2 expresses the proposed C-NOMA-based user pairing in detail. The fact that how frequently the pairing process is executed mainly depends on the URLLC use case. For example, in factory automation with fixed or slow speed devices, the algorithm does not need to perform in each frame. Moreover, it should be noted that these computations are performed at the BS with the assumption of CSIT, and the results are sent to the users.

C. Hybrid pairing

To describe the hybrid pairing, let us first consider the NOMA user pairing scheme proposed in [33]. Pursuant to this, the first strong user is paired with the first weak user; the second strong user is paired with the second weak user, and so on. Accordingly, all the users are paired. The fact is that principle of NOMA is to select users with a high difference in their channel gains. In particular, NOMA's performance diminishes when the difference in channel gains among the users is small. For example, in Fig. 2, user-pairs 6 and 7, which have almost the same channel conditions, may decrease the spectral efficiency and system capacity due to the unsuccessful SIC. Hence, it is sensible that such non-suitable pairs are omitted from NOMA scheduling, and their clustering continues with OMA. In this method, the BS adaptively switches between the NOMA and OMA transmission modes according to the instantaneous strength of wireless channels and hence the performance of the NOMA/OMA user pairing, and each of them that meets the max-min fairness criterion is selected as the access scheme.

Algorithm 2: Joint suboptimal C-NOMA-based user pairing and resource allocation

1 Input: sorted DL channel gains in descending order $|h_1|^2 > |h_2|^2 > \cdots > |h_{2K}|^2$ and the corresponding D2D channel gains, device coverage radius r_0, inputs of Algorithm 1.

2 Output: the user pairing $\mathbf{A} = [a_{i,j}]_{2K \times 2K}$.

3 Determine the graph matrix of the network topology.

4 Set $i := 2K$.

5 while $i \geq 1$ **do**
    // allocating C-NOMA pairs
6   **if** *user i has not been paired* **then**
7   Find the set of unpaired adjacent users of user i, $\boldsymbol{\psi}_i$.
8   **if** $\text{length}(\boldsymbol{\psi}_i) \neq 0$ **then**
9   **for** $l = 1 : \text{length}(\boldsymbol{\psi}_i)$ **do**
10   Calculate $T_0^{i, \boldsymbol{\psi}_i(l)}$ by Algorithm 1.
11   **end**
12   Set $[T_0^*, index] := \max\left\{T_0^{i, \boldsymbol{\psi}_i(l)}\right\}$ and $j := \boldsymbol{\psi}_i(index)$.
13   Pair users i and j, i.e. $a_{i,j} := 1$.
14   **end**
15   **end**
16   Set $i := i - 1$.
17 end
18 Set $i := 2K$ and $j := 1$.
19 while $i > j$ **do**
    // allocating hybrid pairs
20   **if** *user i has not been paired* **then**
21   **while** *user j has been paired* **do**
22   Set $j := j + 1$.
23   **end**
24   Pair users i and j, i.e. $a_{i,j} := 1$.
25   Set $j := j + 1$.
26   **end**
27   Set $i := i - 1$.
28 end
29 Return: $\mathbf{A} = [a_{i,j}]_{2K \times 2K}$.

We discussed the hybrid pairing for scheduling the users that are isolated or left unpaired in the proposed C-NOMA-based user pairing. However, these two basic schemes, namely NOMA and hybrid user pairing, can independently be implemented and are considered as benchmark schemes in our simulations.

VII. NUMERICAL RESULTS

In this section, the proposed C-NOMA scheme's performance along with SC and MRC strategies are evaluated through the numerical results based on our analytical solutions. A heterogeneous network consists of URLLC users with different reliability requirements is considered. PAPR factor and required accuracy in Algorithm 1 are considered as $\kappa_{\mathrm{p}} = 1.2$ and $\epsilon = 10^{-15}$, respectively. Also, it is assumed that $P_{\mathrm{ave}} = 10$ W and $D_{\max} = 200$ channel uses, unless

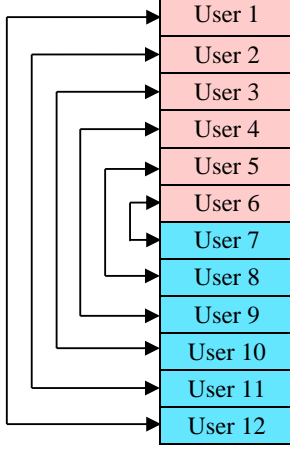Fig. 2. The 2-user NOMA pairing scheme [33].

Fig. 3. Maximum fair throughput achieved by the C-NOMA and NOMA schemes versus $D_{\max}$, when $P_{\mathrm{ave}} = 10$ W.

otherwise stated. The numerical results are provided based on fixed channel gains with two users and random channel gains with more than two users, which are presented in the following two subsections.

A. Two-user Network with Fixed Channel Gains

Throughout this subsection, to provide insight into the relationships between the proposed and the benchmark schemes, the channel gains of the two users are set to be fixed. For instance, it is assumed that $|h_1|^2/\sigma^2 = 0.8$ and $|h_2|^2/\sigma^2 = 0.1$. We investigate the performance of the proposed schemes in two various relaying link status. Meaning, when the two users are near to each other and the relaying link is strong, it is assumed that $|h_{1,2}|^2/\sigma^2 = 0.5$, and when the two users are far from each other, and the relaying link is poor, it is assumed that $|h_{1,2}|^2/\sigma^2 = 0.01$. Meanwhile, users BLER are considered as $\varepsilon_1^{\mathrm{th}} = 10^{-7}$ and $\varepsilon_2^{\mathrm{th}} = 10^{-5}$.

In Fig. 3, the effect of total blocklength, $D_{\max}$, on the fair throughput in the proposed C-NOMA with SC and MRC strategies is assessed in two relaying link modes. Also, the optimal NOMA and OMA results in our previous work [14] are shown for comparison. It is observed that in the strong relaying link mode, both combining strategies applied to the C-NOMA effectively improve the fair throughput compared to the NOMA/OMA. It is also observed that the MRC receiver outperforms the SC receiver, regardless of the blocklength. Because in the combined signal with MRC protocol, SINR increases, so the decoding error probability of user 2 decreases. Hence, it is possible that by less blocklength allocation to phase II, the reliability performance of user 2 can still be guaranteed at the desired level. As a result, more blocklength is allocated to phase I. Hence, users' data rates and system fair throughput increase.

On the other hand, in a poor relaying link, the C-NOMA scheme (in both combining strategies) has exactly the same performance as the NOMA. In fact, in this case, the optimal decision is in favor of the direct link, and the C-NOMA is transformed into the NOMA. However, in a realistic wireless
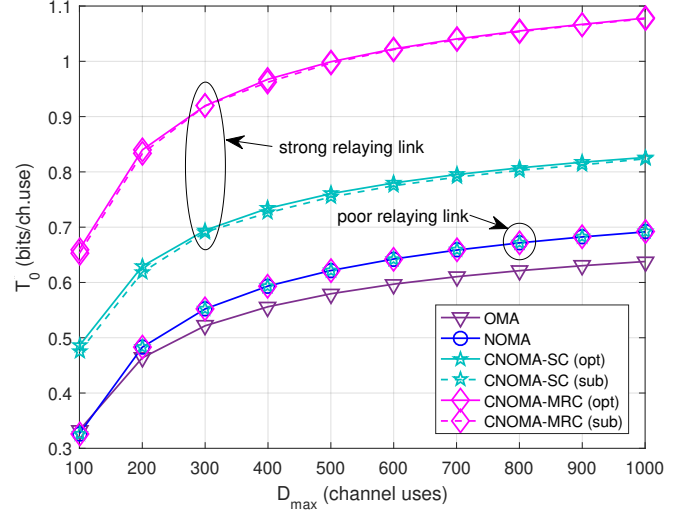
channel, mixed conditions occur together, and C-NOMA outperforms the NOMA on average. Moreover, it is observed that suboptimal solutions in both SC and MRC receivers converge to the near-optimal solutions.

In Fig. 4, the effect of average total power, P_{ave}, on the fair throughput is investigated. In the strong relaying link mode, the C-NOMA's superiority with MRC receiver is notable against the SC receiver and the NOMA/OMA scheme. In addition, the C-NOMA with SC strategy outperforms the NOMA in low power/SNR ranges, while it coincides with the NOMA on average powers greater than 20 W. This could be justified by the fact that in SC strategy, the signals do not combine, and transmission in phase II assures the success of user 2's packet decoding. Hence, in low SNRs where the weak user's probability of successful decoding in phase I is not too high, the reliability is increased by retransmission in phase II. However, in high SNRs, where the allocated power of user 2 in the NOMA phase guarantees the reliability, phase II transmission is pointless. Therefore, in this case, transmission via a single phase is optimal in comparison with two-phase, and the proposed scheme performs like the NOMA. Moreover, in the poor relaying link mode, the C-NOMA scheme always complies with the NOMA. As a result, from the complexity perspective, the C-NOMA usage with SC strategy seems sensible just in low SNR regimes.

B. Multi-user Network with Random Channel Gains

Here, we assume that the BS is located at the center of a cell with radius of 300 m. The system bandwidth is set as $B = 1$ MHz, which is equivalent to a DL transmission duration 0.2 ms for a blocklength of 200 channel uses, and satisfies the low-latency criterion of URLLC standards. The noise power spectral density is -173 dBm/Hz, and small-scale channel coefficients are Rayleigh fading with $\mathcal{CN}(0,1)$ distribution. Large-scale path loss is modeled as $L = 35.3 + 37.6\log_{10}d(\mathrm{m})$
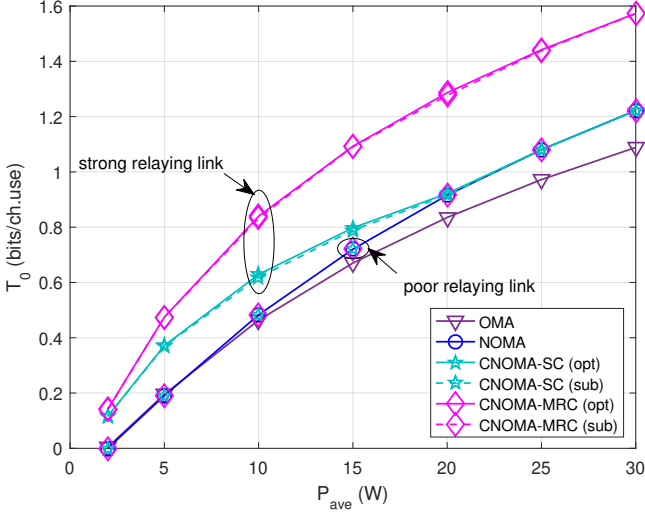
Fig. 4. Maximum fair throughput achieved by the C-NOMA and NOMA schemes versus P_{ave}, when $D_{\max} = 200$.

Fig. 5. Average fair throughput achieved by the different pairing schemes versus the number of users.

dB [24]. The total number of independent channel generations is set as 1000.

Fig. 5 illustrates the average achievable fair throughput versus the number of users, $2K$, for the proposed C-NOMA pairing with SC and MRC strategies. We compare it with exhaustive search method and the method proposed in [26], which is based on pairing one near user and one far user. In that method, first we sort the K^2 D2D channel gains of any near-far pair and delete the $K-1$ weakest channels. Then, by considering the number of deleted channel gains of each far user, first a near user is paired to the far user with the largest-number deleted channel and last the far user with the smallest-number. The selection criteria is to maximize the throughput of the far user. To be comparable with our proposed method, unlike [26], we assume that both near and far users can communicate with the BS directly, and SC/MRC combining schemes are performed at the far user. Moreover, the NOMA and hybrid pairing schemes are illustrated as benchmark. It demonstrates that the proposed C-NOMA pairing scheme (in both MRC and SC techniques) converges to a near-optimal solution. While, the near-far pairing method achieves the lower performance in both combining strategies. On the other hand, the NOMA pairing scheme stated in [33] yields the lowest throughput, especially in the presence of a large number of users, and as expected, the hybrid pairing scheme outperforms the NOMA pairing.

To evaluate the fairness of the proposed C-NOMA-based user pairing, Fig. 6 indicates Jain's fairness index for the proposed scheme and the benchmarks. Jain's fairness index is defined as [34]

$$J = \frac{\left(\sum_{k=1}^{K} T_k^*\right)^2}{K \sum_{k=1}^{K} T_k^{*2}}, \tag{20}$$

where T_k^* indicates the optimal fair throughput of pair k. Jain's fairness index is bounded in $[0, 1]$ which equal users' throughput obtains the maximum value. As Fig. 6 illustrates,
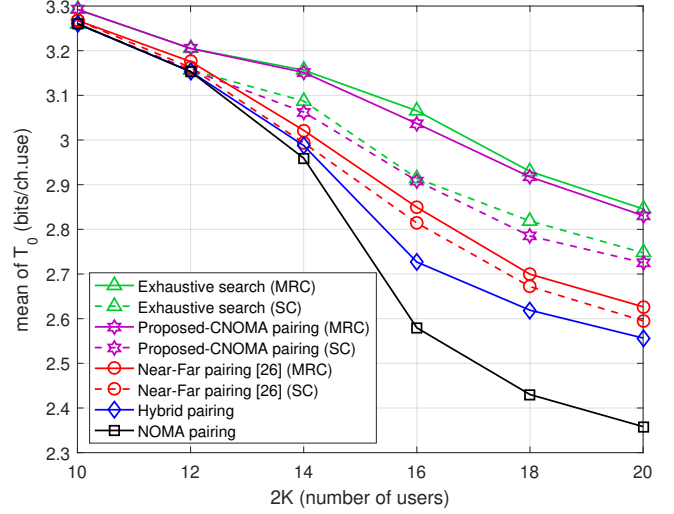
the hybrid pairing scheme is fairer comparing to the C-NOMA-based and the NOMA pairing schemes. The reason is that in the C-NOMA-based pairing schemes, i.e., the proposed, near-far, and exhaustive search methods, creating C-NOMA pairs for all the users is not probable. Hence, unavoidably, some users are scheduled in hybrid NOMA/OMA pairs. Since the C-NOMA users will achieve more throughput than the users with hybrid pairing, the fairness will degrade in these schemes.

Moreover, regarding the logic behind the hybrid pairing, it will always be fairer than the NOMA pairing. Interestingly, the C-NOMA-based pairing schemes (with both combining strategies) result in more fairness relative to the NOMA pairing in the presence of a large number of users. This is due to the fact that the denser the network is, the more users will experience the same channel. This will cause more failures in NOMA scheduling, so the C-NOMA pairing will obtain more fairness in that case.

VIII. CONCLUSION AND FUTURE WORKS

In this paper, the combination of NOMA with the cooperative relaying technique (i.e., C-NOMA) was considered in short packet communications to guarantee high reliability and low latency. The performance of two relaying strategies, i.e., SC and MRC, was presented in terms of decoding error probability in a quasi-static channel. Besides, the necessity to provide QoS of all users with critical services motived us to consider max-min fairness as a design criterion in URLLC systems. To this end, first, an optimization problem was formulated for a two-user DL C-NOMA system, and optimal power, blocklength, and transmission rate were determined under the total energy consumption, reliability, and delay constraints. To decrease the computational complexity, a suboptimal algorithm was proposed with near-optimal performance. Numerical results showed that the proposed C-NOMA scheme improves the users' fair throughput significantly, compared to
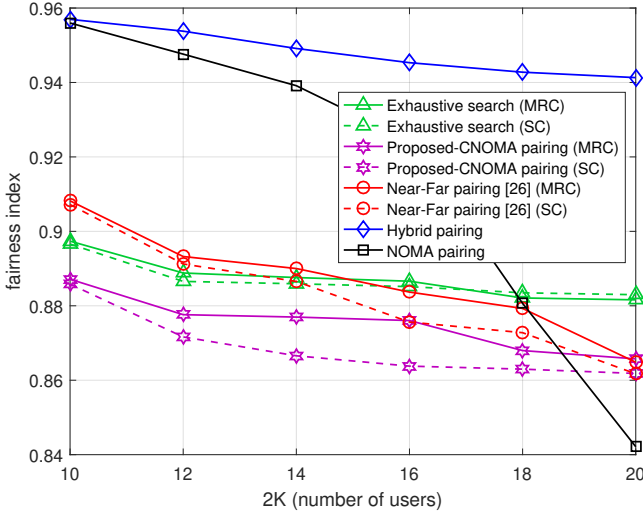
Fig. 6. Fairness comparison between the different pairing schemes versus the number of users.

the NOMA scheme. Moreover, it was demonstrated that the C-NOMA scheme with MRC strategy outperforms SC strategy.

Finally, the problem was extended to a multi-user scenario, and a pairing scheme based on C-NOMA was proposed. Monte Carlo simulations showed that the proposed C-NOMA pairing scheme performs close to the optimal solution, with less computational complexity. Further, the simulation results verify the supremacy of the proposed user pairing (with both SC and MRC techniques) over the near-far pairing method proposed in [26], as well as the NOMA and hybrid OMA/NOMA pairing schemes in boost the average fair throughput despite degrading the fairness index slightly.

The presented work in this paper can be extended from different directions. Two of the main potential extensions that remain for future works are developing the proposed pairing algorithm for the case of statistical CSI, and considering a distributed method for network coordination. The statistical CSI knowledge can remove the shortage of the out-dated CSI and the feedback overhead due to CSIT. Despite distributed method for network coordination might look to be more suitable for URLLC, it requires deep investigation as they normally introduce different types of overhead which may cause additional delay.

APPENDIX A
PROOF OF PROPOSITION 1

We prove Proposition 1 by the contradiction method. We consider the optimal solution of problem (15) as $\left\{p_1^{\mathrm{I}\dagger}, p_2^{\mathrm{I}\dagger}, p_2^{\mathrm{II}\dagger}, m^{\mathrm{I}\dagger}, m^{\mathrm{II}\dagger}\right\}$, where $m^{\mathrm{I}\dagger}(p_1^{\mathrm{I}\dagger} + p_2^{\mathrm{I}\dagger}) < m^{\mathrm{II}\dagger}p_2^{\mathrm{II}\dagger}$. It can achieve the maximum value of $\min\{T_1, T_2\}$, which is denoted by $T_0^\dagger$. We increase $p_1^{\mathrm{I}\dagger}$ and $p_2^{\mathrm{I}\dagger}$ by multiplying in a scalar value $\alpha > 1$ to attain $p_1^{\mathrm{I}^*} = \alpha p_1^{\mathrm{I}\dagger}$ and $p_2^{\mathrm{I}^*} = \alpha p_2^{\mathrm{I}\dagger}$. It can be verified that the following equation holds,

$$m^{\mathrm{I}\dagger}\left(p_1^{\mathrm{I}\dagger} + p_2^{\mathrm{I}\dagger}\right) + m^{\mathrm{II}\dagger}p_2^{\mathrm{II}\dagger} =$$
$$m^{\mathrm{I}\dagger}\left(p_1^{\mathrm{I}^*} + p_2^{\mathrm{I}^*}\right) + m^{\mathrm{II}\dagger}p_2^{\mathrm{II}^*} = D_{\max}P_{\mathrm{ave}}$$

We note that since $\alpha > 1$, so $p_1^{\mathrm{I}^*} > p_1^{\mathrm{I}\dagger}$ and $p_2^{\mathrm{I}^*} > p_2^{\mathrm{I}\dagger}$. Hence, we have $\gamma_{1,1}^{\mathrm{I}^*} > \gamma_{1,1}^{\mathrm{I}\dagger}$ and

$$\gamma_{2,2}^{\mathrm{I}^*} = \frac{p_2^{\mathrm{I}^*}|h_2|^2}{p_1^{\mathrm{I}^*}|h_2|^2 + \sigma^2} = \frac{p_2^{\mathrm{I}\dagger}|h_2|^2}{p_1^{\mathrm{I}\dagger}|h_2|^2 + \frac{\sigma^2}{\alpha}}$$
$$> \frac{p_2^{\mathrm{I}\dagger}|h_2|^2}{p_1^{\mathrm{I}\dagger}|h_2|^2 + \sigma^2} = \gamma_{2,2}^{\mathrm{I}\dagger}$$

This means as $p_i^{\mathrm{I}\dagger}$, $i \in \{1, 2\}$, increases to $p_i^{\mathrm{I}^*}$, the corresponding SNR/SINR increases, which results in an increase in $R_{i,i}^{\mathrm{I}}$ and finally T_i increases (Invoking [14, Appendix A], the allowed $R_{i,i}^{\mathrm{I}}$ is a monotonically increasing function of $\gamma_{i,i}^{\mathrm{I}}$.). On the other hand, $R_{i,i}^{\mathrm{I}}$ and so T_i are clearly increasing functions of m^{I}. Then, we can construct a new solution $\left\{p_1^{\mathrm{I}^*}, p_2^{\mathrm{I}^*}, p_2^{\mathrm{II}^*}, m^{\mathrm{I}^*}, m^{\mathrm{II}^*}\right\}$, where corresponds to T_0^*. Also, $m^{\mathrm{I}^*} = m^{\mathrm{II}\dagger} + \Delta m$ and $m^{\mathrm{II}^*} = m^{\mathrm{II}\dagger} - \Delta m$ with $\Delta m > 0$. As before, this solution satisfies $m^{\mathrm{I}^*}\left(p_1^{\mathrm{I}^*} + p_2^{\mathrm{I}^*}\right) + m^{\mathrm{II}^*}p_2^{\mathrm{II}^*} = D_{\max}P_{\mathrm{ave}}$. Since $p_i^{\mathrm{I}^*} > p_i^{\mathrm{I}\dagger}$ and $m^{\mathrm{I}^*} > m^{\mathrm{I}\dagger}$, we have $T_0^* > T_0^\dagger$. This contradicts the assumption that $T_0^\dagger$ is an optimal solution. So, we can always find a proper α and Δm such that $m^{\mathrm{I}^*}\left(p_1^{\mathrm{I}^*} + p_2^{\mathrm{I}^*}\right) > m^{\mathrm{II}^*}p_2^{\mathrm{II}^*}$.

REFERENCES

[1] F. Salehi, N. Neda, M.-H. Majidi, and H. Ahmadi, "Max-min fairness with selection combining strategy on cooperative noma: A finite block-length analysis," in *2021 Joint European Conference on Networks and Communications 6G Summit (EuCNC/6G Summit)*, 2021, pp. 43–48.

[2] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proceedings of the IEEE*, vol. 104, no. 9, pp. 1711–1726, 2016.

[3] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, 2010.

[4] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static multiple-antenna fading channels at finite blocklength," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 4232–4265, 2014.

[5] P. Wu and N. Jindal, "Coding versus ARQ in fading channels: How reliable should the phy be?" *IEEE Transactions on Communications*, vol. 59, no. 12, pp. 3363–3374, 2011.

[6] B. Makki, T. Svensson, and M. Zorzi, "Finite block-length analysis of the incremental redundancy harq," *IEEE Wireless Communications Letters*, vol. 3, no. 5, pp. 529–532, 2014.

[7] S. Xu, T. Chang, S. Lin, C. Shen, and G. Zhu, "Energy-efficient packet scheduling with finite blocklength codes: Convexity analysis and efficient algorithms," *IEEE Transactions on Wireless Communications*, vol. 15, no. 8, pp. 5527–5540, 2016.

[8] H. Ren, C. Pan, Y. Deng, M. Elkashlan, and A. Nallanathan, "Joint pilot and payload power allocation for massive-mimo-enabled urllc iiot networks," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 5, pp. 816–830, 2020.

[9] ——, "Resource allocation for secure urllc in mission-critical iot scenarios," *IEEE Transactions on Communications*, vol. 68, no. 9, pp. 5793–5807, 2020.

[10] C. She, Y. Duan, G. Zhao, T. Q. S. Quek, Y. Li, and B. Vucetic, "Cross-layer design for mission-critical iot in mobile edge computing systems," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 9360–9374, 2019.

[11] Y. Yu, H. Chen, Y. Li, Z. Ding, and B. Vucetic, "On the performance of non-orthogonal multiple access in short-packet communications," *IEEE Communications Letters*, vol. 22, no. 3, pp. 590–593, 2018.

[12] X. Sun, S. Yan, N. Yang, Z. Ding, C. Shen, and Z. Zhong, "Short-packet downlink transmission with non-orthogonal multiple access," *IEEE Transactions on Wireless Communications*, vol. 17, no. 7, pp. 4550–4564, 2018.

[13] Y. Xu, C. Shen, T. Chang, S. Lin, Y. Zhao, and G. Zhu, "Transmission energy minimization for heterogeneous low-latency noma downlink," *IEEE Transactions on Wireless Communications*, vol. 19, no. 2, pp. 1054–1069, 2020.

[14] F. Salehi, N. Neda, and M.-H. Majidi, "Max-min fairness in downlink non-orthogonal multiple access with short packet communications," *AEU - International Journal of Electronics and Communications*, vol. 114, p. 153028, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1434841119321570

[15] Y. Hu, J. Gross, and A. Schmeink, "On the performance advantage of relaying under the finite blocklength regime," *IEEE Communications Letters*, vol. 19, no. 5, pp. 779–782, 2015.

[16] Y. Hu, A. Schmeink, and J. Gross, "Blocklength-limited performance of relaying under quasi-static rayleigh channels," *IEEE Transactions on Wireless Communications*, vol. 15, no. 7, pp. 4548–4558, 2016.

[17] ——, "Optimal scheduling of reliability-constrained relaying system under outdated csi in the finite blocklength regime," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 7, pp. 6146–6155, 2018.

[18] Z. Ding, M. Peng, and H. V. Poor, "Cooperative non-orthogonal multiple access in 5g systems," *IEEE Communications Letters*, vol. 19, no. 8, pp. 1462–1465, 2015.

[19] Z. Ding, H. Dai, and H. V. Poor, "Relay selection for cooperative noma," *IEEE Wireless Communications Letters*, vol. 5, no. 4, pp. 416–419, 2016.

[20] P. Xu, Y. Wang, G. Chen, G. Pan, and Z. Ding, "Design and evaluation of buffer-aided cooperative noma with direct transmission in iot," *IEEE Internet of Things Journal*, pp. 1–1, 2020.

[21] F. Kara and H. Kaya, "Threshold-based selective cooperative-noma," *IEEE Communications Letters*, vol. 23, no. 7, pp. 1263–1266, 2019.

[22] X. Lai, Q. Zhang, and J. Qin, "Cooperative noma short-packet communications in flat rayleigh fading channels," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 6, pp. 6182–6186, 2019.

[23] Y. Hu, M. C. Gursoy, and A. Schmeink, "Efficient transmission schemes for low-latency networks: Noma vs. relaying," in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, 2017, pp. 1–6.

[24] H. Ren, C. Pan, Y. Deng, M. Elkashlan, and A. Nallanathan, "Joint power and blocklength optimization for urllc in a factory automation scenario," *IEEE Transactions on Wireless Communications*, vol. 19, no. 3, pp. 1786–1801, 2020.

[25] L. Zhu, J. Zhang, Z. Xiao, X. Cao, and D. O. Wu, "Optimal user pairing for downlink non-orthogonal multiple access (NOMA)," *IEEE Wireless Communications Letters*, vol. 8, no. 2, pp. 328–331, 2019.

[26] Y. Cheng, K. H. Li, K. C. Teh, S. Luo, and W. Wang, "Two-step user pairing for OFDM-based cooperative NOMA systems," *IEEE Communications Letters*, vol. 24, no. 4, pp. 903–906, 2020.

[27] J. Zhang, X. Tao, H. Wu, and X. Zhang, "Performance analysis of user pairing in cooperative NOMA networks," *IEEE Access*, vol. 6, pp. 74 288–74 302, 2018.

[28] P. Hũu, M. A. Arfaoui, S. Sharafeddine, C. M. Assi, and 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*, vol. 68, no. 11, pp. 6737–6749, 2020.

[29] A. K. Lamba, R. Kumar, and S. Sharma, "Joint user pairing, subchannel assignment and power allocation in cooperative non-orthogonal multiple access networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 10, pp. 11 790–11 799, 2020.

[30] C. She, C. Yang, and T. Q. S. Quek, "Joint uplink and downlink resource configuration for ultra-reliable and low-latency communications," *IEEE Transactions on Communications*, vol. 66, no. 5, pp. 2266–2280, 2018.

[31] G. J. Sutton, J. Zeng, R. P. Liu, W. Ni, D. N. Nguyen, B. A. Jayawickrama, X. Huang, M. Abolhasan, Z. Zhang, E. Dutkiewicz, and T. Lv, "Enabling technologies for ultra-reliable and low latency communications: From phy and mac layer perspectives," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2488–2524, 2019.

[32] S. Timotheou and I. Krikidis, "Fairness for non-orthogonal multiple access in 5g systems," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1647–1651, 2015.

[33] M. S. Ali, H. Tabassum, and E. Hossain, "Dynamic user clustering and power allocation for uplink and downlink non-orthogonal multiple access (NOMA) systems," *IEEE Access*, vol. 4, pp. 6325–6343, 2016.

[34] R. Jain, D.-M. Chiu, and W. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems," *CoRR*, vol. cs.NI/9809099, 1998. [Online]. Available: https://arxiv.org/abs/cs/9809099

Fateme Salehi received the B.Sc. and M.Sc. degrees in communication engineering from University of Birjand (Birjand, Iran) in 2010 and 2012 respectively. Currently she is a Ph.D. student in communication engineering in Department of Electrical and Computer Eng. at the University of Birjand, Iran. Since March 2021, she has joined KTH Royal Institute of Technology as a visiting researcher. Her research interests include signal processing, channel estimation, MIMO-OFDM and NOMA systems, resource management in URLLC, and the Internet-of-Things.

Naaser Neda received the B.S. degree in electrical Eng. from the University of Tehran and the M.S. degree in communication Eng. from Sharif University of Technology (SUT), both in Tehran, Iran, in 1990 and 1994 respectively. He received the PhD degree in communication Eng. from the University of Surrey (CCSR), Guildford, UK, in 2003. He is currently an Associate Professor of communication engineering with the Department of Electrical and Computer Eng. at the University of Birjand, Iran. His research interests include signal processing for communication systems, physical layer of CDMA/MCCDMA/OFDM networks, sensor networks, and NOMA.

Mohammad-Hassan Majidi received the B.S. degree in electrical Eng. from the Shahid Bahonar University of Kerman and the M.S. degree in communication Eng. from Imam Hossein Comprehensive University of Tehran, both in Iran, in 2003 and 2006 respectively. He received the PhD degree in telecommunication engineering from the Department of Telecommunications, Ecole Superieure d'Electricite (Supelec), Gif-sur-Yvette, France, in 2013. He is currently an Assistant Professor of communication engineering with the Department of Electrical and Computer Eng. at the University of Birjand, Iran. His research interests include signal processing for communication systems, joint channel and data detection, cryptography and secure communication.

Hamed Ahmadi received the Ph.D. degree from the National University of Singapore, in 2012. He was an Agency for Science Technology and Research (A-STAR) funded Ph.D. student at the Institute for Infocomm Research (I2R), National University of Singapore. Since then, he has been working with different academic and industrial positions in Ireland and U.K. He is currently an Assistant Professor with the Department of Electronic Engineering, University of York, U.K. He is also an Adjunct Assistant Professor with the school of Electrical and Electronic Engineering, University College Dublin, Ireland. He has published more than 50 peer-reviewed book chapters, journal, and conference papers. His current research interests include design, analysis, and optimization of wireless communications networks, airborne networks, wireless network virtualization, blockchain, the Internet-of-Things, cognitive radio networks, and the application of machine learning in small cell and self-organizing networks. He is a member of Editorial Board of IEEE ACCESS, Frontiers in Blockchain, and Wireless Networks (Springer). He is a Fellow of the U.K., Higher Education Academy and Networks working Group Co-Chair and a management committee member of COST Action 15104 (IRACON).