Network-Coded NOMA With Antenna Selection for the Support of Two Heterogeneous Groups of Users

Amjad Saeed Khan, *Member, IEEE*, Ioannis Chatzigeorgiou, *Senior Member, IEEE*, Sangarapillai Lambotharan, *Senior Member, IEEE*, and Gan Zheng, *Senior Member, IEEE*

***Abstract*— The combination of non-orthogonal multiple access (NOMA) and transmit antenna selection (TAS) techniques has recently attracted significant attention due to the low cost, low complexity, and high diversity gains. Meanwhile, random linear coding (RLC) is considered to be a promising technique for achieving high reliability and low latency in multicast communications. In this paper, we consider a downlink system with a multi-antenna base station and two multicast groups of single-antenna users, where one group can afford to be served opportunistically, while the other group consists of comparatively low-power devices with limited processing capabilities that have strict quality of service (QoS) requirements. In order to boost reliability and satisfy the QoS requirements of the multicast groups, we propose a cross-layer framework, including NOMA-based TAS at the physical layer and RLC at the application layer. In particular, two low-complexity TAS protocols for NOMA are studied in order to exploit the diversity gain and meet the QoS requirements. In addition, RLC analysis aims to facilitate heterogeneous users, such that sliding window-based sparse RLC is employed for computational restricted users, and conventional RLC is considered for others. Theoretical expressions that characterize the performance of the proposed framework are derived and verified through simulation results.**

***Index Terms*— Antenna selection, non-orthogonal multiple access, network coding, random block matrices, outage probability, decoding probability, throughput.**

I. Introduction

DUE to the explosive increase in demand for wireless connectivity, both the academic and industrial communities have increased their research focus on the design of fifth generation (5G) wireless networks. 5G networks are envisioned to support very high data rates, extremely low latency, a manifold increase in base station capacity and high quality of service (QoS) [1]. To this end, Non-Orthogonal Multiple Access (NOMA) has been recognized as a promising multiple access technique for next generation communications, and has attracted considerable research attention [2]–[4]. NOMA can significantly improve the spectral efficiency and user fairness of mobile communication networks. For example, NOMA exploits the power domain to allocate more power to users experiencing weaker channel conditions in order to guarantee user fairness [5], and allows multiple users to simultaneously share the same resource block (e.g. frequency, time or code). Thus, it greatly improves network capacity [6] as compared to schemes based on orthogonal multiple access [7]. Undoubtedly, NOMA is regarded as one of the key technologies for supporting massive connectivity, dense coverage and low latency in 5G.

Besides NOMA, network reliability and capacity can be improved by employing multiple-input multiple-output (MIMO) technology [8], [9]. However, the performance gains from the use of multiple antennas come at the cost of increased computational complexity and power consumption that scale with the number of antennas [10]. In order to avoid the undesirable effects that arise from the simultaneous use of multiple antennas but, at the same time, preserve the diversity and throughput benefits, antenna selection (AS) has been recognized as a practical solution in the literature [11]. As demonstrated in [12], AS techniques can achieve full diversity gain. Recently, AS in combination with NOMA has attracted significant attention [13]–[16]. For example, the outage performance for downlink NOMA was investigated in [13] by employing Transmit AS (TAS) at the base station. In addition, efficient AS techniques were proposed in [14] and [15] to maximize the sum-rate in downlink MIMO-NOMA networks.

Given that NOMA is considered to be a promising access scheme for 5G networks [2], it should be in a position to support a variety of services and devices, from real-time video for laptop and smartphone users to low-rate data for computationally bounded and energy constrained sensors. In the case of multicast services, only a limited number of retransmissions is allowed, so that resources are not wasted in an effort to accommodate every user of a multicast group who was not successful in recovering the transmitted data. In this context, Random Linear Coding (RLC) – also referred to as *randomized network coding* [17] – has gained popularity as a key technology that can improve network throughput,

Manuscript received June 20, 2018; revised October 22, 2018; accepted January 1, 2019. Date of publication January 16, 2019; date of current version February 11, 2019. This work was supported in part by the Engineering and Physical Sciences Research Council under Grant EP/R006385/1 and Grant EP/N007840/1, and in part by the Leverhulme Trust Research Project under Grant RPG-2017-129. The associate editor coordinating the review of this paper and approving it for publication was X. Yuan. *(Corresponding author: Amjad Saeed Khan.)*

A. S. Khan, S. Lambotharan, and G. Zheng are with the Wolfson School of Mechanical, Electrical and Manufacturing Engineering, Loughborough University, Loughborough LE11 3TU, U.K. (e-mail: a.khan@lboro.ac.uk; s.lambotharan@lboro.ac.uk; g.zheng@lboro.ac.uk).

I. Chatzigeorgiou is with the School of Computing and Communications, Lancaster University, Lancaster LA1 4WA, U.K. (e-mail: i.chatzigeorgiou@lancaster.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2019.2891769

robustness [18] and latency [19] by reducing the number of distinct transmissions. A novel characteristic of RLC from other end-to-end information combining paradigms is the requirement of intermediate network nodes to linearly combine incoming source or coded packets, generate new coded packets and forward them to the next set of nodes. Thus, RLC can achieve the multicast capacity [18], [20]. Moreover, the inherent properties of RLC make it a suitable candidate for cases where the objective is not only to improve network reliability but also energy consumption [21], routing complexity [22], storage efficiency [23] and security [24].

RLC can also facilitate the co-existence of heterogeneous devices with different processing power, size and storage limitations. In order to accommodate a diverse set of receiving devices, the data that are about to be transmitted by a base station or an access point can be divided into priority layers, which are encoded using RLC that offers Unequal Error Protection (UEP) [25], [26]. The priority layers usually consist of a base layer and multiple enhancement layers. The base layer is responsible for providing a basic QoS, suitable for devices with small storage and limited processing power. The enhancement layers contain data that can further improve the QoS. Thus, a high QoS will be offered to a device that can access all or as many layers as possible. This layered structure of RLC fits well into various applications. For example, it has been considered in [27] as Prioritized Random Linear Coding (PRLC) for layered data delivery from multiple servers, in [28] as UEP RLC for wireless layered video broadcasting and in [29] as Expanding Window-RLC for multimedia multicast services based on H.264/SVC.

A limitation of RLC is its decoding complexity. For instance, in order to decode K source packets, each of size S symbols from a given finite field, the decoder needs to perform $O(K^3 + K^2\ S)$ finite field operations to invert a $K \times K$ matrix using the Gaussian elimination algorithm [30]. Practical methods that aim to reduce the decoding complexity of RLC include the adoption of chunk codes [31], the implementation of RLC over non-overlapping windows [26] and the use of RLC over disjoint generations [32]. These schemes first split a message into disjoint sub-messages. The packets of each sub-message compose a generation, also known as a window. Each sub-message is then encoded separately using RLC. The decoding complexity, which is inversely proportional to the number of partitioned sub-messages, is lower than that of conventional RLC over the whole message. However, this reduction in complexity comes at the cost of reduced performance (in terms of decoding probability) and increased overhead (in terms of transmitted coded packets). In an effort to fine-tune the trade-off between the performance advantage of conventional RLC and the reduced decoding complexity of RLC based on disjoint generations, the partitioned sub-messages can be allowed to overlap. This RLC implementation is known as overlapping generations [32], overlapped chunk codes [33] and sliding window RLC [34], [35]. The aforementioned schemes exploit a principle similar to that of message passing, used by fountain decoders [36]; packets of decoded generations can be back-substituted into undecoded generations that contain them and increase the probability of decoding these generations, thus improving the overall throughput. In order to further reduce both the decoding complexity and the overhead while maintaining the delay performance, sparse RLC within each generation as well as a feedback mechanism to control the amount of overlap between generations were proposed in [30] and [37].

The introduction of RLC in NOMA was first explored in [38]. Recently, opportunistic RLC was proposed in [39] to improve the reliability and diversity gain in cooperative NOMA. The key objective of this paper is to tailor TAS to downlink NOMA that employs different RLC schemes to accommodate multicast groups having different service requirements and computational capabilities. For instance, devices with strong processing capabilities could support RLC over large finite fields but computationally bounded devices could require the implementation of reduced decoding complexity RLC. The main contributions of the paper can be summarized as follows:

- Exact theoretical expressions for the decoding probability and the average decoding delay of RLC based on sliding windows are derived. Similar expressions are not available in the literature and both metrics are usually computed through simulations, e.g., [34].
- Low complexity TAS protocols for NOMA are presented and analytical expressions for their outage probabilities are obtained. Outage probabilities are equivalent to frame error rates in quasi-static fading channels, if the signal-to-noise ratio (SNR) threshold in the outage analysis reflects the modulation and coding scheme at the physical layer.
- The proposed RLC-enabled NOMA-based TAS scheme is investigated and extensive simulations are conducted to validate the accuracy of the derived expressions. A comparison between the NOMA-based scheme and its OMA-based counterpart, in terms of required transmission power, decoding probability and network throughput, is provided.

The rest of this paper is organized as follows: Section II describes the system model, provides definitions and introduces relevant notations. Section III presents RLC schemes suitable for heterogeneous users and obtains theoretical expressions for the decoding probability of each coding scheme. A detailed description of NOMA-based TAS protocols is provided in Section IV. Exact closed-form expressions for evaluating the network performance and throughput are also derived. Results are discussed in Section V and conclusions are drawn in Section VI.

II. System Model

We consider a downlink scenario where a base station (BS) broadcasts to two multicast groups of users $\mathcal{U}_1$ and $\mathcal{U}_2$ as shown in Fig. 1. The BS is equipped with a set of N_A antennas $\mathcal{N} = \{1, 2, \ldots, N_{\mathrm{A}}\}$, which it uses to perform transmit antenna selection in order to maximize system throughput. On the other hand, users in the two groups are equipped with a single antenna, which performs in half duplex mode. We assume that $\mathcal{U}_2$ is associated to low-rate delay-sensitive applications, while $\mathcal{U}_1$ is related to high-rate delay-tolerant applications that can afford opportunistic connectivity.

Moreover, we assume that all users of group $\mathcal{U}_2$ are comparatively low-power devices with limited processing capabilities. For example, $\mathcal{U}_1$ could be a group of users associated with multimedia applications, while $\mathcal{U}_2$ could be a group of Internet of Things (IoT) sensors or healthcare devices. All links are assumed to be quasi-static Rayleigh fading channels. The BS broadcasts data to all users in both groups but ceases transmission only if designated coordinators from each group, that is, $\mathrm{U}_1 \in \mathcal{U}_1$ and $\mathrm{U}_2 \in \mathcal{U}_2$, successfully receive and decode the data intended for their respective groups. A coordinator can locally communicate with users in its own group and is responsible for accommodating requests by other users for additional information after the BS has ceased transmission. Hence, in contrast to other users in groups $\mathcal{U}_1$ and $\mathcal{U}_2$, the successful delivery of information to both U_1 and U_2 is of high priority for the BS, and will thus be the focus of the rest of this paper. Note that, the number of BS transmissions can also be controlled by the coordinator selection. For example, the user with the worst channel condition in a group can be chosen as a coordinator, such that instead of relying on the coordinator's support, the BS can be forced to transmit enough coded packets to ensure the successful decoding by all the users of the group. In other words, for delay sensitive groups a coordinator selection can be tuned such that users can be served at once by the BS, without any delay caused by the coordinator's assistance. On the other hand, the user with the best channel conditions can be a coordinator of a delay tolerant group. Various methods for the selection of appropriate coordinators based on the channel state information (CSI) of users have been discussed in [40]. However, the selection process of coordinators is not within the scope of this paper.

The gains h_{i1} and h_{i2} of the channels between the i^{th} antenna of the BS and the two coordinators of the multicast groups have been modeled as zero-mean, independent but not identically distributed complex Gaussian random variables with variances σ_1^2 and σ_2^2, respectively. The quasi-static nature of the channel implies that the value of h_{iu} remains constant for the duration of a transmitted frame of symbols but changes independently from frame to frame. For the transmission of each frame, the BS employs one of the following TAS protocols in order to select an appropriate antenna:

1) *Conventional TAS*: The transmit antenna that provides the best channel quality between the BS and U_2 is selected.
2) *Two stage TAS*: Transmit antennas that strictly satisfy the QoS requirements of U_2 are identified in the first stage. Among them, the best antenna for U_1 is selected in the second stage.

Note that, two stage TAS aims to provide better user fairness [14] under the condition that the QoS requirement of U_2 is satisfied. On the other hand, conventional TAS is considered as a suitable technique when the BS only knows the channel state information of U_2. In addition, it is presented as a benchmark protocol in order to compare the performance gain of the proposed two stage TAS. The total transmit power at the BS is limited to P. The fraction of the power that BS allocates to the symbols meant for multicast group $\mathcal{U}_u$ is a_u, where $u \in \{1, 2\}$, so that $a_1 + a_2 = 1$. Let $\gamma_{u,\nu,i} = \rho a_\nu |h_{iu}|^2$ denote the instantaneous SNR of a signal transmitted by the i^{th} antenna of the BS, meant for group $\mathcal{U}_\nu$ and received by the user coordinator in group $\mathcal{U}_u$, where $\rho = P/N_0$ with N_0 as the variance of the additive white Gaussian noise.

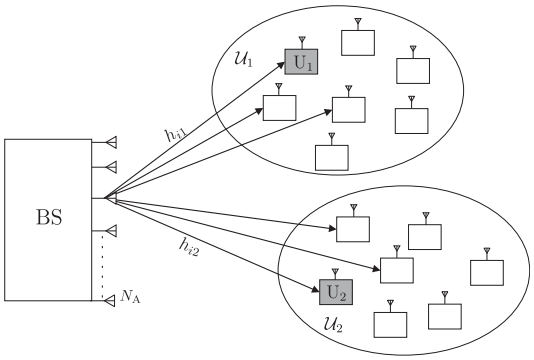

Fig. 1. Block diagram of the network depicting the base station BS and the multicast groups $\mathcal{U}_1$ and $\mathcal{U}_2$, with the selected coordinators U_1 and U_2.

As shown in [41], the modulation and coding scheme (MCS) employed by U_u can be characterised by an SNR threshold, denoted by $\hat{\gamma}_u$, if the channel gain follows the Rayleigh distribution. Using this threshold-based approach, the frame error rate of a point-to-point system can be tightly approximated by the outage probability. If the channel gain follows the more general Nakagami distribution, the threshold-based approach can still be used and the SNR threshold can be determined as shown in [42]–[44]. We assume that the channel conditions between BS and U_1 are better than those between BS and U_2, i.e., $|h_{i1}|^2 > |h_{i2}|^2$. We also assume that U_1 employs an MCS that offers higher throughput but suffers from higher sensitivity to channel errors than the MCS of U_2. Consequently, the relationship between the SNR thresholds of the two users is $\hat{\gamma}_1 > \hat{\gamma}_2$.

Let the multicast group $\mathcal{U}_1$ request the transmission of a data file of size K_1 source packets, and let $\mathcal{U}_2$ expect to receive a data file of size K_2 source packets, as depicted in Fig. 2. We denote by $\mathbf{s}_{u,j} \in \{0,1\}^{S_u}$ the j-th source packet meant for $\mathcal{U}_u$, where $j \in \{1, \ldots, K_u\}$ and S_u is the size of a source packet in bits. In order to boost the network reliability and throughput, the BS employs RLC at the application layer and randomly combines the K_u source packets $\{\mathbf{s}_{u,1}, \ldots, \mathbf{s}_{u,K_u}\}$ in order to generate $K'_u \geq K_u$ coded packets $\{\mathbf{c}_{u,1}, \ldots, \mathbf{c}_{u,K'_u}\}$.

The number of coded packets K'_u can be potentially infinite but is usually set to a value that reflects the time and power constraints of the system. A coded packet $\mathbf{c}_{u,j}$ can be obtained as follows

$$\mathbf{c}_{u,j} = \sum_{k=1}^{K_u} g_{u,j,k}\, \mathbf{s}_{u,k} \tag{1}$$

where arithmetic operations are over a Galois field of q_u elements, denoted by $\mathbb{F}_{q_u}$. Depending on the value of j and k, the adopted RLC scheme sets the coding coefficient $g_{u,j,k}$ to a value that is selected independently and at random from $\mathbb{F}_{q_u}$. Coded packets and source packets have the same length S_u but a coding vector that contains the coding

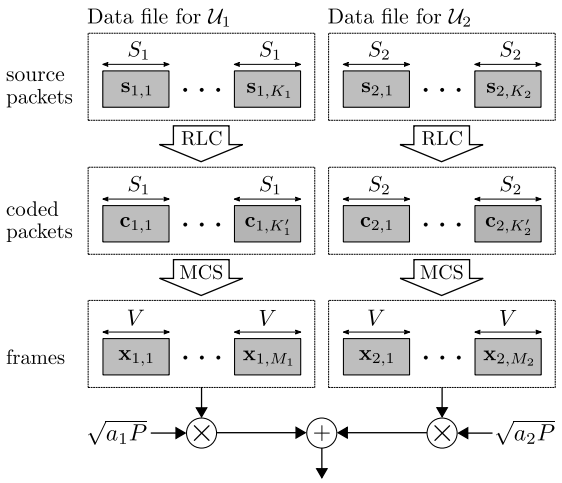

Fig. 2. The BS employs RLC to segment and encode the data files for multicast groups $\mathcal{U}_1$ and $\mathcal{U}_2$, and uses the NOMA protocol to superimpose the generated frames.

coefficients $\{g_{u,j,1}, \ldots, g_{u,j,K_u}\}$ is customarily appended to $\mathbf{c}_{u,j}$ in order to assist the receiver in the reconstruction of the source packets. In practice, the coding vector can be represented by the seed value of a predetermined pseudo-random function [45] or shortened using simple compression methods [46] before it is appended to the header of the associated coded packet. Therefore, the overhead introduced by the coding vectors is negligible compared to the size of a data file, and is not considered in the remainder of the paper.

At the physical layer, the MCS converts the stream of K_u' coded packets, each of which is composed of S_u bits, into a sequence of M_u frames, each of which contains the same fixed number of V symbols. After the addition of headers and padding by the intermediate network layers, we assume that a frame at the physical layer carries $\tau_u \in \mathbb{N}^+$ coded packets,[1] i.e., $M_u = K_u'/\tau_u$. Let $\mathbf{x}_{u,m}$ denote the m-th frame to be delivered to group $\mathcal{U}_u$. The two frames $\mathbf{x}_{1,m}$ and $\mathbf{x}_{2,m}$ are superimposed using NOMA, on a symbol-by-symbol basis, and form the transmitted frame $\mathbf{x}_m = \sqrt{a_1}\mathbf{x}_{1,m} + \sqrt{a_2}\mathbf{x}_{2,m}$. The frame received by user U_u can be represented as:

$$\mathbf{y}_{u,m,i^*} = \sqrt{P} h_{i^*u} \left(\sqrt{a_1}\,\mathbf{x}_{1,m} + \sqrt{a_2}\,\mathbf{x}_{2,m}\right) + \mathbf{w} \qquad (2)$$

where i^* specifies the antenna selected by the adopted TAS protocol, and $\mathbf{w}$ is a vector of V circularly-symmetric Gaussian random variables with zero mean and variance N_0. Note that $a_2 \geq a_1$ in order to satisfy the QoS requirements of users of group $\mathcal{U}_2$. In this paper, we consider fixed values of power allocation coefficients a_1 and a_2. Optimization of the power allocation coefficients is beyond the scope of this paper. When U_1 receives a superimposed frame, it employs successive interference cancellation to extract and subtract frame $\mathbf{x}_{2,m}$ from $\mathbf{y}_{u,m,i^*}$. The desired frame $\mathbf{x}_{1,m}$ will be recovered if $\gamma_{1,1,i^*} \geq \hat{\gamma}_1$. User U_2 is expected to extract $\mathbf{x}_{2,m}$ without performing interference cancellation because $\gamma_{2,2,i^*} > \gamma_{2,1,i^*}$. The desired frame $\mathbf{x}_{2,m}$ will be recovered if $\gamma_{2,2,i^*}/(\gamma_{2,1,i^*}+1) \geq \hat{\gamma}_2$. The values of the thresholds $\hat{\gamma}_1$ and $\hat{\gamma}_2$ depend both on the MCS employed by each multicast group and the length V of the transmitted frames. User U_u will successfully reconstruct the K_u source packets if K_u linearly independent coded packets are obtained from the recovered frames.

We remark that, for $1 \leq m \leq \min(M_1, M_2)$, the transmitted frame $\mathbf{x}_m$ is the result of the superposition of frames $\mathbf{x}_{1,m}$ and $\mathbf{x}_{2,m}$. When $\min(M_1, M_2) < m \leq \max(M_1, M_2)$, the frames of one of the multicast groups have been delivered and the BS would allocate all its available power to transmit the remaining frames to the other multicast group. In this paper, we consider the worst-case scenario according to which the BS always has data for both multicast groups in the queue and, consequently, the transmitted frame is always the superposition of two frames.

The key parameters of the system model have been summarized in Table I for quick reference. The following sections explain the considered RLC schemes and derive expressions for the probability of a user recovering a transmitted data file.

TABLE I
KEY PARAMETERS OF THE SYSTEM MODEL

Notation	Description
BS	Transmitting base station
$\mathcal{U}_u$	Receiving multicast group u, where $u \in \{1, 2\}$
U_u	Representative user in group $\mathcal{U}_u$
K_u	Number of source packets to be encoded for $\mathcal{U}_u$
K_u'	Number of coded packets to be transmitted to $\mathcal{U}_u$
M_u	Number of frames that carry the K_u' coded packets
S_u	Length (in bits) of a source or coded packet for $\mathcal{U}_u$
V	Length (in symbols) of a frame for $\mathcal{U}_1$ or $\mathcal{U}_2$
$\mathbf{x}_{\nu,m}$	The m-th frame at BS containing coded packets for $\mathcal{U}_\nu$
$\mathbf{y}_{u,m,i^*}$	The m-th noisy frame transmitted by antenna i^* and received by U_u
ρ	Set to P/N_0, where P is the available transmit power at BS and N_0 is the variance of AWGN
a_ν	Proportion of P allocated for the transmission of $\mathbf{x}_{\nu,m}$
h_{iu}	Fading coefficient of the channel between i^{th} antenna of the BS and U_u
$\gamma_{u,\nu,i}$	Instantaneous SNR of a signal for U_ν at U_u transmitted by i^{th} antenna of the BS
$\hat{\gamma}_u$	Required SNR threshold for signal recovery at U_u

III. RLC SCHEMES FOR HETEROGENEOUS USERS

The coded packets in a frame that has been successfully recovered by the physical layer of user U_u are forwarded to the application layer. Given that the RLC decoding process is the same for both multicast groups, we drop the index u from the notation in this section. With this in mind, let N denote the number of coded packets that have been delivered to the application layer of user U. The K coding coefficients associated to each coded packet are stacked to form a $N \times K$ decoding matrix $\mathbf{D}$. The K source packets

[1] In this paper, $\mathbb{N}^+$ denotes the set of all natural numbers excluding zero, i.e., $\mathbb{N}^+ = \{1, 2, \ldots\}$, while $\mathbb{N}_0 = \mathbb{N}^+ \cup \{0\}$.

will be retrieved and the data file will be reconstructed if K linearly independent coded packets are collected. This implies that matrix $\mathbf{D}$ has full rank and, therefore, contains a $K \times K$ invertible matrix.

A. Classic RLC

In classic RLC (c-RLC), the value of each coding coefficient is selected uniformly at random from $\mathbb{F}_q$. For $N \geq K$, the number of all full-rank realizations of the $N \times K$ matrix $\mathbf{D}$ is given by [47, p. 338]

$$f(N,K) = \begin{cases} \prod_{i=0}^{K-1} (q^N - q^i) & \text{if } K \geq 1 \\ 1, & \text{if } K = 0 \end{cases} \tag{3}$$

while the number of all $N \times K$ matrix realizations having a specific rank r, for $0 \leq r \leq \min(N,K)$, is equal to

$$f_r(N,K) = \frac{f(N,r)\, f(K,r)}{f(r,r)} \tag{4}$$

as explained in [48] and [49]. If $\mathbb{F}_q^{N\times K}$ denotes the set of all $N \times K$ matrices over $\mathbb{F}_q$, the probability that a particular realization of matrix $\mathbf{D}$ has full rank can be obtained by dividing $f(N,K)$ by q^{NK}, which is the cardinality of $\mathbb{F}_q^{N\times K}$. More specifically, we obtain

$$P_{\mathrm{c}}(N,K) = \frac{f(N,K)}{q^{NK}} = \prod_{i=0}^{K-1} (1 - q^{-N+i}). \tag{5}$$

This well-known expression establishes that the larger the size of the Galois field is, the higher the probability of recovering the data file is. However, RLC over large Galois fields incurs a significant decoding cost, in terms of memory footprint, computational complexity and energy requirements [50]. To alleviate this problem, computationally bounded and energy constrained devices could resort to RLC over Galois fields of a small size, e.g., $\mathbb{F}_2$. Alternatively, they could employ *sparse* RLC over large Galois fields.

In sparse RLC, the number of source packets involved in the generation of each coded packet is kept small or, equivalently, a coding coefficient is more likely to be zero than in classic RLC. As reported in [51] and demonstrated in [52], systems employing sparse RLC feature a significantly reduced decoding complexity than classic RLC but require a larger number of transmitted coded packets in order to recover the source packets. A popular implementation of sparse RLC is *sliding window* RLC, according to which only source packets that have indices within a moving range of values are randomly selected and contribute to the generation of coded packets. The following subsection focuses on the sliding window RLC and derives expressions for the probability of recovering a data file when the sliding window RLC is used.

B. Sliding Window RLC

The use of a sliding window mechanism for the selection of a subset of source packets, based on which coded packets are generated, was proposed in [34] for random fountain codes and extended to Raptor codes in [53]. The concept of a window sliding over the source packets was later introduced into RLC for wireless mesh networks [35] and networks compatible with the Transmission Control Protocol (TCP) [54]. Sliding window mechanisms are also being considered by the Network Coding Research Group of the Internet Research Task Force (IRTF) for the practical implementation of network coding in future Internet architectures [55].

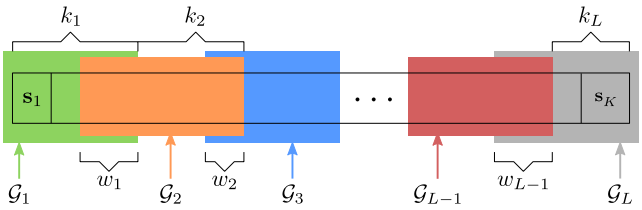

Fig. 3. Example of sw-RLC. The K source packets of the data file are members of L overlapping generations $\mathcal{G}_1, \ldots, \mathcal{G}_L$. For $i > 1$, generations $\mathcal{G}_{i-1}$ and $\mathcal{G}_i$ have w_{i-1} source packets in common.

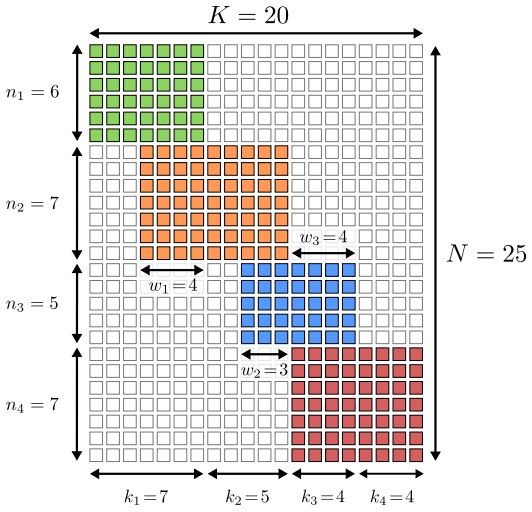

Fig. 4. Example of a 25×20 decoding matrix at a user when the BS employs sw-RLC. The source packets have been divided into $L = 4$ generations. Random elements are depicted by colored solid squares (■), while zero-valued entries are represented by empty squares (□).

In a similar fashion to other practical network coding schemes, sliding window RLC (sw-RLC) organizes the K source packets into L groups, referred to as *generations* [56]. Let the i-th generation, denoted by $\mathcal{G}_i$, contain ℓ_i source packets. Fig. 3 shows the implementation of sw-RLC that we consider in this paper. Observe that the L generations overlap, such that generation $\mathcal{G}_i$ shares w_{i-1} of its ℓ_i source packets with generation $\mathcal{G}_{i-1}$ only, that is, $|\mathcal{G}_{i-1} \cap \mathcal{G}_i| = w_{i-1}$. If k_i is the number of source packets in $\mathcal{G}_i$ that are not shared with $\mathcal{G}_{i-1}$, we can write $\ell_i = w_{i-1} + k_i$, where $\ell_1 = k_1$ for $i = 1$, while $\sum_{i=1}^{L} k_i = K$. The number of shared packets between generations $\mathcal{G}_{i-1}$ and $\mathcal{G}_i$ can take values in the range $0 \leq w_{i-1} \leq k_{i-1}$.

The RLC encoder at the BS generates k_i' coded packets from generation $\mathcal{G}_i$, and user U recovers n_i of those packets, for $i = 1, \ldots, L$, with $\sum_{i=1}^{L} n_i = N$. Fig. 4 gives an example of the structure of the $N \times K$ decoding matrix $\mathbf{D}$ when sw-RLC is employed. In this example, $K = 20$ source packets

have been grouped into $L = 4$ overlapping generations with sizes $\ell_1 = 7$, $\ell_2 = 9$, $\ell_3 = 7$ and $\ell_4 = 8$. Adjacent generations share $w_1 = 4$, $w_2 = 3$ and $w_3 = 4$ source packets, as shown in Fig. 4. User U has obtained $N = 25$ coded packets and will recover all of the source packets if and only if the rank of matrix $\mathbf{D}$ in this example is 20.

We shall refer to random matrices conforming to the general structure of decoding matrices generated by sw-RLC as *block tri-diagonal* (BTD) matrices. BTD matrices can be defined by a 3-tuple of row vectors $\{\mathbf{n}, \mathbf{k}, \boldsymbol{\ell}\}$, where $\mathbf{n} = [n_1, \ldots, n_L]$ contains the number of received coded packets associated to each generation, $\mathbf{k} = [k_1, \ldots, k_L]$ dictates the number of source packets that a generation does not have in common with the previous generation, and $\boldsymbol{\ell} = [\ell_1, \ldots, \ell_L]$ contains the number of source packets in each generation. An expression for enumerating full-rank BTD matrices is not readily available in the literature and is derived in the remainder of this section. Before we proceed with the proof of one lemma, which will lead us to the main proposition, we first introduce some additional notation. If $\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_L$ are matrices having the same number of columns, then $(\boldsymbol{\Phi}_1; \ldots; \boldsymbol{\Phi}_L)$ denotes the matrix obtained by the *vertical concatenation* of the L matrices or, equivalently, by appending $\boldsymbol{\Phi}_{i+1}$ to the bottom of $\boldsymbol{\Phi}_i$, for $i = 1, \ldots, L-1$. In a similar fashion, we denote by $(\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_L)$ the *horizontal concatenation* of L matrices, provided that they all have the same number of rows.

Lemma 1: Let $\mathbf{D} = (\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2) \in \mathbb{F}_q^{N \times K}$ *be a random BTD matrix that has the following structure:*

$$\mathbf{D} = \left[\begin{array}{cc|c} \overbrace{\boldsymbol{\Phi}_1^{(1)}}^{k_1 - w_1} & \overbrace{\boldsymbol{\Phi}_1^{(2)}}^{w_1} & \overbrace{\boldsymbol{\Phi}_2}^{k_2} \\ \mathbf{0} & \boldsymbol{\Phi}_1^{(3)} & \boldsymbol{\Phi}_2^{(1)} \end{array}\right] \begin{matrix} \} n_1 \\ \} n_2 \end{matrix}$$

where $\boldsymbol{\Phi}_1 \in \mathbb{F}_q^{N \times k_1}$, $\boldsymbol{\Phi}_2 \in \mathbb{F}_q^{N \times k_2}$, $k_1 + k_2 = K$, $n_1 + n_2 = N$ *and* $N \geq K$. *The number of full-rank realizations of matrix* $\mathbf{D}$ *can be expressed as*

$$f_{\mathrm{BTD}}(\mathbf{n}, \mathbf{k}, \boldsymbol{\ell}) = f(n_1, k_1 - w_1)\, f(n_2, k_2) \prod_{i=K-w_1}^{K-1} \left(q^N - q^i\right) \tag{6}$$

or, equivalently,

$$f_{\mathrm{BTD}}(\mathbf{n}, \mathbf{k}, \boldsymbol{\ell}) = \sum_{r_1} \prod_{i=1}^{2} f_{r_i}(n_{i+1}, w_i) f(n_i - r_{i-1}, k_i - r_i) q^{\varphi_i} \tag{7}$$

where $\max(0, k_i - n_i + r_{i-1}) \leq r_i \leq \min(n_{i+1} - k_{i+1}, w_i)$ *and* $\varphi_i = (k_i - r_i) w_{i-1} + n_i r_i$, *while* $n_3 = 0$, $w_0 = w_2 = 0$ *and* $r_0 = r_2 = 0$.

Proof: For matrix $\mathbf{D}$ to be full-rank, its K columns should be linearly independent. The first $k_1 - w_1$ columns and the last k_2 columns of matrix $\mathbf{D}$ will be linearly independent, if sub-matrix $\boldsymbol{\Phi}_1^{(1)}$ is chosen from the pool of $f(n_1, k_1 - w_1)$ matrices that have rank $k_1 - w_1$, and sub-matrix $\boldsymbol{\Phi}_2^{(1)}$ is one of the $f(n_2, k_2)$ matrix realizations that have rank k_2. Matrix $\mathbf{D}$ will have rank K if the w_1 columns of $(\boldsymbol{\Phi}_1^{(2)}; \boldsymbol{\Phi}_1^{(3)})$ are also linearly independent of each other and, at the same time, linearly independent of the columns of $\boldsymbol{\Phi}_1^{(1)}$ and $\boldsymbol{\Phi}_2^{(1)}$. There are $q^N - q^{k_1 + k_2 - w_1}$ choices for the first column of the $N \times w_1$ matrix $(\boldsymbol{\Phi}_1^{(2)}; \boldsymbol{\Phi}_1^{(3)})$ to be linearly independent of the previously considered $k_1 + k_2 - w_1$ columns of $\mathbf{D}$. Following this reasoning for every column of $(\boldsymbol{\Phi}_1^{(2)}; \boldsymbol{\Phi}_1^{(3)})$ and recalling that $k_1 + k_2 = K$, we conclude that the total number of $N \times w_1$ full-rank matrices that could be part of the $N \times K$ full-rank matrix $\mathbf{D}$ is $(q^N - q^{K - w_1}) \cdot \ldots \cdot (q^N - q^{K-1})$. Based on the aforementioned arguments, the number of full-rank realizations of matrix $\mathbf{D}$ is given by (6).

Alternatively, we could start by assuming that $\boldsymbol{\Phi}_1^{(3)}$ is one of the $f_{r_1}(n_2, w_1)$ matrices that have rank r_1. If the rank of $\boldsymbol{\Phi}_1^{(3)}$ is r_1, $\boldsymbol{\Phi}_1$ also contains at least r_1 linearly independent columns. For $\boldsymbol{\Phi}_1$ to have rank k_1 and be a full-rank matrix, its remaining $k_1 - r_1$ columns should also be linearly independent. This can only happen if the corresponding $k_1 - r_1$ columns of the $n_1 \times k_1$ matrix $(\boldsymbol{\Phi}_1^{(1)}, \boldsymbol{\Phi}_1^{(2)})$ are linearly independent and form one of the $f(n_1, k_1 - r_1)$ matrices that have rank $k_1 - r_1$. Given that there are $q^{n_1 r_1}$ choices for the remaining r_1 columns of $(\boldsymbol{\Phi}_1^{(1)}, \boldsymbol{\Phi}_1^{(2)})$, we conclude that the number of full-rank realizations of $\boldsymbol{\Phi}_1$ is

$$f_{\boldsymbol{\Phi}_1} = f_{r_1}(n_2, w_1)\, f(n_1, k_1 - r_1)\, q^{n_1 r_1}. \tag{8}$$

On the other hand, $\boldsymbol{\Phi}_2$ will have rank k_2 if $\boldsymbol{\Phi}_2^{(1)}$ consists of k_2 linearly independent rows that are also independent of the r_1 rows of $\boldsymbol{\Phi}_1^{(3)}$. The total number of full-rank matrices $\boldsymbol{\Phi}_2$ is given by

$$f_{\boldsymbol{\Phi}_2} = f(n_2 - r_1, k_2)\, q^{w_1 k_2} \tag{9}$$

and the number of full-rank realizations of $\mathbf{D}$ can be obtained by summing the product $f_{\boldsymbol{\Phi}_1} f_{\boldsymbol{\Phi}_2}$ over all valid values of r_1, that is

$$f_{\mathrm{BTD}}(\mathbf{n}, \mathbf{k}, \boldsymbol{\ell}) = \sum_{r_1} f_{r_1}(n_2, w_1) f(n_1, k_1 - r_1) q^{n_1 r_1} \cdot f(n_2 - r_1, k_2)\, q^{w_1 k_2}. \tag{10}$$

Expression (10) can take the form of (7) if we define $n_3 = 0$, $w_0 = w_2 = 0$, $r_0 = r_2 = 0$, $\varphi_i = (k_i - r_i) w_{i-1} + n_i r_i$ and $f_0(0, 0) = 1$. The equivalence of expressions (6) and (7) is also proven analytically in the Appendix. ■

Both (6) and (7) generate the same output values for the same input parameters. Even though relationship (6) is more elegant, relationship (7) captures the correlation of overlapping generations and can be extended to any number of generations, as the following proposition demonstrates.

Proposition 1: Let $\mathbf{D} \in \mathbb{F}_q^{N \times K}$ *be a random BTD matrix that has the following structure:*

$$\mathbf{D} = \left[\begin{array}{cccc|ccc} \boldsymbol{\Phi}_1^{(1)} & \boldsymbol{\Phi}_1^{(2)} & 0 & 0 & \cdots & \cdots & 0 \\ 0 & \boldsymbol{\Phi}_1^{(3)} & \boldsymbol{\Phi}_2^{(1)} & \boldsymbol{\Phi}_2^{(2)} & \cdots & \cdots & 0 \\ 0 & 0 & 0 & \boldsymbol{\Phi}_2^{(3)} & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \cdots & \boldsymbol{\Phi}_L^{(1)} \end{array}\right]$$

with $\mathbf{\Phi}_i^{(1)} \in \mathbb{F}_q^{n_i \times (k_i - w_i)}$, $\mathbf{\Phi}_i^{(2)} \in \mathbb{F}_q^{n_i \times w_i}$ *and* $\mathbf{\Phi}_i^{(3)} \in \mathbb{F}_q^{n_{i+1} \times w_i}$ *for* $i = 1, \ldots, L-1$, *and* $\mathbf{\Phi}_L^{(1)} \in \mathbb{F}_q^{n_L \times k_L}$. *Furthermore,* $\sum_{i=1}^{L} k_i = K$, $\sum_{i=1}^{L} n_i + n_2 = N$ *and* $N \geq K$. *The number of full-rank realizations of matrix* $\mathbf{D}$ *is given by*

$$f_{\mathrm{BTD}}(\mathbf{n}, \mathbf{k}, \boldsymbol{\ell}) = \sum_{r_1} \cdots \sum_{r_{L-1}} \prod_{i=1}^{L} f_{r_i}(n_{i+1}, w_i) \times f(n_i - r_{i-1}, k_i - r_i) q^{\varphi_i} \quad (11)$$

where $\max(0, k_i - n_i + r_{i-1}) \leq r_i \leq \min(n_{i+1}, w_i)$ *and* $\varphi_i = (k_i - r_i) w_{i-1} + n_i r_i$, *while* $n_{L+1} = 0$, $w_0 = w_L = 0$ *and* $r_0 = r_L = 0$.

Proof: Let $\mathbf{\Phi}_i$ denote the $N \times k_i$ submatrix of $\mathbf{D}$ that contains $\mathbf{\Phi}_i^{(1)}$, $\mathbf{\Phi}_i^{(2)}$ and $\mathbf{\Phi}_i^{(3)}$ for $i = 1, \ldots, L-1$, and let $\mathbf{\Phi}_L$ denote the $N \times k_L$ submatrix of $\mathbf{D}$ that contains $\mathbf{\Phi}_L^{(1)}$. Furthermore, let r_i be the rank of $\mathbf{\Phi}_i^{(3)}$. Matrix $\mathbf{D}$ can be seen as the horizontal concatenation of L matrices, that is, $\mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_L$. Matrix $\mathbf{\Phi}_1$ can be generated in $f_{\mathbf{\Phi}_1}$ different ways, as explained in Lemma 1. Recall that $\mathbf{\Phi}_1$ is only affected by the value of r_1, which allows for $k_1 - r_1$ columns of $(\mathbf{\Phi}_1^{(1)}, \mathbf{\Phi}_2^{(1)})$ to form a full-rank submatrix. The number of ways that $\mathbf{\Phi}_2$ can be constructed depends on the values of both r_1 and r_2. In particular, the number of columns that can contribute to a full-rank submatrix in $(\mathbf{\Phi}_2^{(1)}, \mathbf{\Phi}_2^{(2)})$ reduces to $k_2 - r_2$ but the number of rows also reduces to $n_2 - r_1$ to ensure independence of the r_1 rows of the adjacent matrix $\mathbf{\Phi}_1^{(3)}$. Taking into account the number of ways that $\mathbf{\Phi}_2^{(3)}$ can be constructed so that it has rank r_2, and considering the number of choices for the remaining elements in $\mathbf{\Phi}_2$, we obtain

$$f_{\mathbf{\Phi}_2} = f_{r_2}(n_3, w_2)\, f(n_2 - r_1, k_2 - r_2)\, q^{(k_2 - r_2) w_1 + n_2 r_2}.$$

The same line of thought can be used to find that

$$f_{\mathbf{\Phi}_i} = f_{r_i}(n_{i+1}, w_i)\, f(n_i - r_{i-1}, k_i - r_i)\, q^{(k_i - r_i) w_{i-1} + n_i r_i}$$

enumerates the ways of obtaining $\mathbf{\Phi}_i$, for $i = 2, \ldots, L-1$. Similarly to (9) in Lemma 1, we find that $\mathbf{\Phi}_L$ can be constructed in $f_{\mathbf{\Phi}_L}$ ways, where

$$f_{\mathbf{\Phi}_L} = f(n_L - r_{L-1}, k_L)\, q^{w_{L-1} k_L}.$$

Summing the product $f_{\mathbf{\Phi}_1} f_{\mathbf{\Phi}_2} \ldots f_{\mathbf{\Phi}_L}$ over all valid values of $r_1, \ldots, r_{L-1}$ gives (11). ∎

Following the same reasoning as in (3) in Section III-A, the probability of obtaining a full-rank realization of the decoding matrix $\mathbf{D}$, when the BS employs sw-RLC, can be obtained as follows

$$P_{\mathrm{sw}}(\mathbf{n}, \mathbf{k}, \boldsymbol{\ell}) = \frac{f_{\mathrm{BTD}}(\mathbf{n}, \mathbf{k}, \boldsymbol{\ell})}{q^{\mathbf{n} \cdot \boldsymbol{\ell}^{\mathsf{T}}}} \quad (12)$$

where $\boldsymbol{\ell}^{\mathsf{T}}$ denotes the transpose of $\boldsymbol{\ell}$, and $\mathbf{n} \cdot \boldsymbol{\ell}^{\mathsf{T}}$ enumerates the elements of the BTD matrix $\mathbf{D}$ that take values from $\mathbb{F}_q$.

Thus, having in mind the system model presented in Section II, we assume that c-RLC is used to offer a reliable and high-rate service to users in group $\mathcal{U}_1$. On the other hand, sw-RLC is employed for the transmission of low-rate data to computationally bounded users of group $\mathcal{U}_2$. In both cases, arithmetic operations are over the same Galois field $\mathbb{F}_q$.

This section has presented expressions (5) and (12) for the probability of recovering all of the K source packets and, hence, being able to reconstruct the transmitted data file, when $N \geq K$ of the K' coded packets have been successfully delivered to a user. The following section considers the complete RLC-enabled NOMA system, computes the overall probability of a user retrieving the desired data file, and evaluates the system throughput.

IV. Performance Analysis

Derivation of the probability of user U_u recovering the K_u source packets of a data file requires knowledge of the frame error rate ε_u, that is, the probability that a frame containing τ_u coded packets will be received in error by user U_u. As explained in Section I, the frame error rate can be tightly approximated by the outage probability, if the SNR threshold $\hat{\gamma}_u$ that characterizes the MCS scheme of user U_u is utilized. In this section, we present antenna selection protocols in greater detail and characterize their performance in terms of the outage probability at both U_1 and U_2. Furthermore, we evaluate the network performance in terms of probability of data recovery and system throughput.

A. Antenna Selection Protocols and Outage Probabilities

1) Conventional TAS: This protocol only considers the channel quality of the link between the BS and U_2. According to this protocol, the BS selects an antenna which provides the maximum channel gain for U_2, given as

$$i^* = arg \max_{i \in \mathcal{N}} \ (h_{i2}). \quad (13)$$

Calculation of the frame error rate at user U_2 is straightforward given that successive interference cancellation is not used. The link between the BS and U_2 will be in outage if the power of the desired signal at U_2 is not higher than the power of the interfering signal and the added noise. Based on [41], we can write

$$\varepsilon_2 = \Pr\left(\frac{\gamma_{2,2,i^*}}{(\gamma_{2,1,i^*} + 1)} \leq \hat{\gamma_2}\right). \quad (14)$$

Given that all channels are statistically independent, the principle of ordered statistics in [57] yields:

$$\begin{aligned} \varepsilon_2 &= \prod_{i=1}^{N_{\mathrm{A}}} \Pr\left(h_{i2} \leq \frac{\hat{\gamma}_2}{\rho(a_2 - a_1 \hat{\gamma}_2)}\right) \\ &= \prod_{i=1}^{N_{\mathrm{A}}} 1 - \exp\left(-\frac{\hat{\gamma}_2}{\sigma_2^2 \rho(a_2 - a_1 \hat{\gamma}_2)}\right). \end{aligned} \quad (15)$$

The link between the BS and user U_1 will be in outage if successive interference cancellation is not successful or the interfering signal is successfully removed but the desired signal cannot be recovered. The frame error rate can thus be expressed as follows:

$$\begin{aligned} \varepsilon_1 &= 1 - \Pr\left(\frac{\gamma_{1,2,i^*}}{(\gamma_{1,1,i^*} + 1)} \geq \hat{\gamma}_2, \ \gamma_{1,1,i^*} \geq \hat{\gamma_1}\right) \\ &= 1 - \Pr\{h_{i^*1} \geq \frac{\hat{\gamma}_2}{\rho(a_2 - a_1 \hat{\gamma}_2)}, \ h_{i^*1} \geq \frac{\hat{\gamma}_1}{\rho a_1}\}. \end{aligned}$$

At U_1's channel, the strongest transmit antenna for U_2 corresponds to a random transmit antenna for U_1, therefore we obtain:

$$\varepsilon_1 = 1 - \exp(-\max(\frac{\hat{\gamma}_2}{\sigma_1^2\rho(a_2 - a_1\hat{\gamma}_2)}, \frac{\hat{\gamma}_1}{\sigma_1^2\rho a_1})). \quad (16)$$

It is important to note that, expressions (15) and (16) present a constraint on the power allocation coefficients, that is, successful delivery requires $a_2 > a_1\hat{\gamma}_2$. For the asymptotic expressions of the outage probabilities, we employ the approximation $1 - \exp(-x) \simeq x$ for $x \to 0$. Thus, for sufficiently high SNR, i.e., $\rho \to \infty$, (15) and (16) can be expressed as:

$$\varepsilon_2^{\infty} \simeq \frac{1}{\rho^{N_A}} \left[\frac{\hat{\gamma}_2}{\sigma_2^2(a_2 - a_1\hat{\gamma}_2)}\right]^{N_A} \quad (17)$$

$$\varepsilon_1^{\infty} \simeq \frac{1}{\rho} \max(\frac{\hat{\gamma}_2}{\sigma_1^2(a_2 - a_1\hat{\gamma}_2)}, \frac{\hat{\gamma}_1}{\sigma_1^2 a_1}). \quad (18)$$

The above expressions imply that, the conventional TAS achieves a diversity gain of order N_A at U_2, however, a diversity gain of order 1 only can be achieved at U_1. This protocol is considered as a benchmark protocol in order to evaluate the performance gain of the proposed two-stage TAS protocol.

2) Two-Stage TAS Protocol: In order to satisfy the QoS requirements of U_2 while being fair to U_1 and thus improve the overall network performance, this protocol considers the channel quality of both the links connecting BS to U_2 and U_1. According to this protocol, In stage I, the BS selects a set of antennas $\mathcal{S}_\mathrm{r}$ which can satisfy the QoS of U_2. In stage II, from the set of selected antennas $\mathcal{S}_\mathrm{r}$, the BS selects the best antenna for U_1. As implied in (15), the constraint on the channel quality $h_{i2} > \frac{\hat{\gamma}_2}{\rho(a_2 - a_1\hat{\gamma}_2)}$ needs to be met for successful delivery to U_2. Therefore, the set $\mathcal{S}_\mathrm{r}$ can be defined as

$$\mathcal{S}_\mathrm{r} = \left\{1 \le i \le N_\mathrm{A} : h_{i2} > \frac{\hat{\gamma}_2}{\rho(a_2 - a_1\hat{\gamma}_2)}\right\}. \quad (19)$$

According to the protocol, only the best antenna in $\mathcal{S}_\mathrm{r}$ is selected to serve U_1, which can be expressed as:

$$i^* = arg\max_{i \in \mathcal{S}_\mathrm{r}} (h_{i1}). \quad (20)$$

Based on the operation of the two-stage TAS protocol, the outage probability at U_2 can be defined as $\varepsilon_2 = \Pr(|\mathcal{S}_\mathrm{r}| = 0)$, where $|\mathcal{S}_\mathrm{r}|$ specifies the number of antennas in $\mathcal{S}_\mathrm{r}$. Thus, the closed form expression of ε_2 can be obtained using (15) as in conventional TAS. On the other hand, the outage probability at U_1 can be obtained from:

$$\varepsilon_1 = \Pr(|\mathcal{S}_\mathrm{r}| > 0)\left[1 - \Pr(h_{i^*1} \ge \frac{\hat{\gamma}_2}{\rho(a_2 - a_1\hat{\gamma}_2)}, h_{i^*1} \ge \frac{\hat{\gamma}_1}{\rho a_1})\right] + \varepsilon_2.$$

In order to derive a closed form expression for ε_1, let $P_{i,\mathrm{out}}$ represent the outage probability due to the selection of the i^{th} antenna in $\mathcal{S}_\mathrm{r}$, given as:

$$\begin{aligned} P_{i,\mathrm{out}} &= 1 - \Pr\{h_{i1} \ge \frac{\hat{\gamma}_2}{\rho(a_2 - a_1\hat{\gamma}_2)}, h_{i1} \ge \frac{\hat{\gamma}_1}{\rho a_1}\} \\ &= 1 - \exp(-\max(\frac{\hat{\gamma}_2}{\sigma_1^2\rho(a_2 - a_1\hat{\gamma}_2)}, \frac{\hat{\gamma}_1}{\sigma_1^2\rho a_1})). \end{aligned} \quad (21)$$

Using ordered statistics and invoking the law of total probability, the expression for ε_1 can be rewritten as:

$$\varepsilon_1 = \varepsilon_2 + \sum_{l=1}^{N_\mathrm{A}} \Pr(|\mathcal{S}_\mathrm{r}| = l) \prod_{i=1}^{l} P_{i,\mathrm{out}} \quad (22)$$

where

$$\begin{aligned} \Pr(|\mathcal{S}_\mathrm{r}| = l) = \binom{N_\mathrm{A}}{l} \prod_{j=1}^{l} \exp(-\frac{\hat{\gamma}_2}{\sigma_2^2\rho(a_2 - a_1\hat{\gamma}_2)}) \\ \times \prod_{k=l+1}^{N_\mathrm{A}} 1 - \exp(-\frac{\hat{\gamma}_2}{\sigma_2^2\rho(a_2 - a_1\hat{\gamma}_2)}). \end{aligned} \quad (23)$$

By exploiting the exponential approximation when $\rho \to \infty$, the asymptotic expression of ε_1^{∞} can be obtained as:

$$\begin{aligned} \varepsilon_1^{\infty} \simeq \frac{1}{\rho^{N_\mathrm{A}}} \Bigg[\Big\{\frac{\hat{\gamma}_2}{\sigma_2^2(a_2 - a_1\hat{\gamma}_2)}\Big\}^{N_\mathrm{A}} + \sum_{l=1}^{N_\mathrm{A}} \binom{N_\mathrm{A}}{l} \\ \cdot\Big\{\frac{\hat{\gamma}_2}{\sigma_2^2(a_2 - a_1\hat{\gamma}_2)}\Big\}^{N_\mathrm{A}-l} \Big\{\max(\frac{\hat{\gamma}_2}{\sigma_1^2(a_2 - a_1\hat{\gamma}_2)}, \frac{\hat{\gamma}_1}{\sigma_1^2 a_1})\Big\}^{l}\Bigg]. \end{aligned}$$

The asymptotic expression of ε_2^{∞} can be obtained using (17). Thus, the derived asymptotic expressions establish that two-stage TAS can provide the full diversity gain of order N_A to both U_1 and U_2.

B. Probability of Data Recovery and Throughput

As discussed in Section III, c-RLC has been used to encode the K_1 source packets of the data file to be broadcast to $\mathcal{U}_1$, and sw-RLC has been used to encode the K_2 source packets of the data file to be broadcast to $\mathcal{U}_2$. Therefore, in order to obtain the overall probability of user U_1 retrieving the data file, expression (5) needs to be averaged over all possible combinations of received frames. In particular, let $B(m, M_u, \varepsilon_u)$ denote the probability mass function of the binomial distribution, given by

$$B(m, M_u, \varepsilon_u) = \binom{M_u}{m} (1 - \varepsilon_u)^m \varepsilon_u^{M_u - m}. \quad (24)$$

The probability of user U_1 recovering the K_1 source packets can be expressed as

$$P_{\mathrm{c},1}^{\mathrm{tot}}(M_1, K_1, \varepsilon_1, \tau_1) = \sum_{m=M_1^{\min}}^{M_1} B(m, M_1, \varepsilon_1)\, P_\mathrm{c}(m\tau_1, K_1) \quad (25)$$

where $m\tau_1$ is the number of coded packets contained in the m frames received by user U_1, $M_1^{\min} = \lfloor (K_1 - 1)/\tau_1 + 1 \rfloor$ is the minimum number of frames that need to be received by user U_1 before the decoding algorithm attempts to recover the K_1 source packets, and $\lfloor x \rfloor$ denotes the integer part of x. The average number of frame transmissions required by U_1 to recover the entire data file can be evaluated using [58] as follows

$$E_{\mathrm{U}_1}(M_1) = M_1 - \sum_{v=0}^{D_1 - 1} P_{\mathrm{c},1}^{\mathrm{tot}}(M_1^{\min} + v, K_1, \varepsilon_1, \tau_1) \quad (26)$$

where D_1 represents the maximum permissible number of excess frame transmissions, that is, $D_1 = M_1 - M_1^{\min}$.

Given that sw-RLC is employed for group $\mathcal{U}_2$, the M_2 transmitted frames contain coded packets associated with generations $\mathcal{G}_1, \ldots, \mathcal{G}_L$. Let $\mathbf{M}_2 = [M_{1,2}, \ldots, M_{L,2}]$ be a row vector, where $M_{i,2}$ denotes the number of frames that carry coded packets associated with generation $\mathcal{G}_i$, for $i = 1, \ldots, L$ and $\sum_{i=1}^{L} M_{i,2} = M_2$. Similarly, let $\mathbf{m} = [m_1, \ldots, m_L]$ be a row vector, with m_i denoting the number of frames that user U_2 has received out of the $M_{i,2}$ transmitted frames, for $i = 1, \ldots, L$. If $\mathbf{m}$ belongs to a set $\mathcal{S}$ defined as follows

$$\mathcal{S} = \{\mathbf{m} \in \mathbb{N}_0^L \mid M_2^{\min} \leq \sum_{i=1}^{L} m_i \leq M_2 \text{ and } m_i \leq M_{i,2}\}.$$

where $M_2^{\min} = \lfloor (K_2 - 1)/\tau_2 + 1 \rfloor$. The probability that user U_2 will retrieve the K_2 source packets admits the form

$$P_{\mathrm{sw},2}^{\mathrm{tot}}(\mathbf{M}_2, \mathbf{k}, \boldsymbol{\ell}, \varepsilon_2, \tau_2) = \sum_{\mathbf{m} \in \mathcal{S}} \prod_{i=1}^{L} B(m_i, M_{i,2}, \varepsilon_2) \cdot P_{\mathrm{sw}}(\mathbf{m}\tau_2, \mathbf{k}, \boldsymbol{\ell}) \quad (27)$$

where $\mathbf{m}\tau_2$ is a row vector conveying the number of received coded packets per generation, and $\mathbf{k}$ and $\boldsymbol{\ell}$ are row vectors that describe the sw-RLC process, as explained in Section III-B. The event of user U_1 decoding the K_1 source packets is indepenent of the event of user U_2 decoding the K_2 source packets. For this reason, the probability that both users will decode their respective source packets, referred to as *joint decoding probability* hereafter, can be obtained from $P_{\mathrm{joint}} = P_{\mathrm{c},1}^{\mathrm{tot}} \times P_{\mathrm{sw},2}^{\mathrm{tot}}$. Note that group $\mathcal{U}_i$ is in outage if the corresponding coordinator U_i is not able to decode the entire source packets. This implies that the outage of group $\mathcal{U}_i$ is associated with the decoding failure of coordinator U_i. Thus, the system is in outage when the coordinators of both groups fail to recover the source packets, or equivalently, $P_{\mathrm{out}} = 1 - P_{\mathrm{joint}}$.

In order to evaluate the average number of frame transmissions required by U_2 to recover the K_2 source packets, we consider implementations that have the following design characteristics:

1) Contiguous generations always share the same number of source packets, i.e., $w_1 = \ldots = w_{L-1} = \hat{w}$.
2) The number of source packets that a generation does not share with the preceding generation is always the same, i.e., $k_1 = \ldots = k_L = \hat{k}$. Hence, the total number of source packets can be expressed as $K_2 = \hat{k}L$.
3) The number of frames transmitted for each generation of source packets is given as $M_{1,2} = \ldots = M_{L-1,2} = \hat{M}$, where $\hat{M} = \lfloor \frac{\hat{k}-1}{\tau_2} + 1 \rfloor$. As $w_L = 0$, we set $M_{L,2} = \hat{M} + \delta$, where $0 \leq \delta \leq \delta_{\max}$ with $\delta_{\max} = \lfloor \frac{\hat{w}-1}{\tau_2} + 1 \rfloor$, to give the last generation a fair chance to be recovered.

Similarly to (26), $E_{\mathrm{U}_2}(M_2)$ denotes the average number of frame transmissions that are required by user U_2 to recover the K_2 source packets, provided that the number of transmitted frames will not exceed M_2. Let $M_2 = \delta_{\max} + (\hat{M} + D_{\mathrm{G}})L$, where D_{G} is the maximum number of excess frame transmissions per generation. To derive an expression for $E_{\mathrm{U}_2}(M_2)$, the increase of frame transmissions from $K_2 = \hat{M}L$ to M_2 can be decomposed into two steps. In the first step, the increase from $\hat{M}L$ to $\delta_{\max} + \hat{M}L$ is considered. The expected number of excess frames that need to be transmitted in support of generation $\mathcal{G}_L$, so that all generations are recovered, can be written as

$$\Delta_{\delta_{\max}} = \delta_{\max} - \sum_{\delta=0}^{\delta_{\max}-1} P_{\mathrm{sw},2}^{\mathrm{tot}}(\mathbf{M}_\delta, \mathbf{k}, \boldsymbol{\ell}, \varepsilon_2, \tau_2) \quad (28)$$

where the first $L-1$ entries of the $1 \times L$ vector $\mathbf{M}_\delta$ are equal to $\hat{M}$ and its last entry is set to $\hat{M} + \delta$. In the second step, the focus is on the increase of the transmitted frames from $\delta_{\max} + \hat{M}L$ to $\delta_{\max} + (\hat{M} + D_{\mathrm{G}})L$. The expected number of excess frame transmissions per generation, denoted by $\Delta_{D_{\mathrm{G}}}$, can be obtained as follows

$$\Delta_{D_{\mathrm{G}}} = D_{\mathrm{G}} - \sum_{v=0}^{D_{\mathrm{G}}-1} P_{\mathrm{sw},2}^{\mathrm{tot}}(\mathbf{M}_v, \mathbf{k}, \boldsymbol{\ell}, \varepsilon_2, \tau_2) \quad (29)$$

where the first $L-1$ entries of the $1 \times L$ vector $\mathbf{M}_\delta$ are equal to $\hat{M} + v$ and its last entry is set to $\hat{M} + \delta_{\max} + v$. Based on (28) and (29), we obtain

$$E_{\mathrm{U}_2}(M_2) = M_2^{\min} + \Delta_{\delta_{\max}} + L\Delta_{D_{\mathrm{G}}} \quad (30)$$

where, $M_2^{\min} = \lfloor \frac{K_2-1}{\tau_2} + 1 \rfloor$. As discussed in Section II, the successful delivery of the appropriate data file to the coordinator of a group guarantees that all users in that group will obtain copies of the data file. Based on the definition of the end-to-end throughput in [59] and [39], the average throughput of the considered network can be defined as:

$$\eta = \frac{\max(M_1^{\min}, M_2^{\min})}{\max(E_{\mathrm{U}_1}(M_1), E_{\mathrm{U}_2}(M_2))}. \quad (31)$$

The dependence of the average throughput on the number of transmit antennas at the BS and the channel coding scheme at the physical layer will be investigated in the following section.

V. Results and Discussion

This section presents simulation results and compares them with analytical results in order to validate the accuracy of the derived expressions. In addition, the performance of NOMA-based TAS combined with RLC schemes is discussed and compared with the performance of a conventional OFDMA-based implementation, which will be referred as OMA. Even though transmissions to the two different groups will not interfere with each other in OMA, the same QoS (in terms of rate) as in NOMA can only be offered if the SNR thresholds are increased. This will lead to an increase in the erasure probabilities ε_1 and ε_2, a reduction in the probability of data recovery, as per (25) and (27), and an increase in the number of frame transmissions, as per (26), (28), (29), and (30).

A Monte Carlo simulation platform representing the system model was developed in MATLAB. Instances, where user coordinator U_u successfully recovers the K_u source packets, were counted and averaged over 10^6 realizations to compute the decoding probability, for $u = 1, 2$. The BS and the users U_1 and U_2 have been positioned such that $\sigma_1^2 = 2.9155$ and $\sigma_2^2 = 0.1715$. We set fixed values to the power allocation

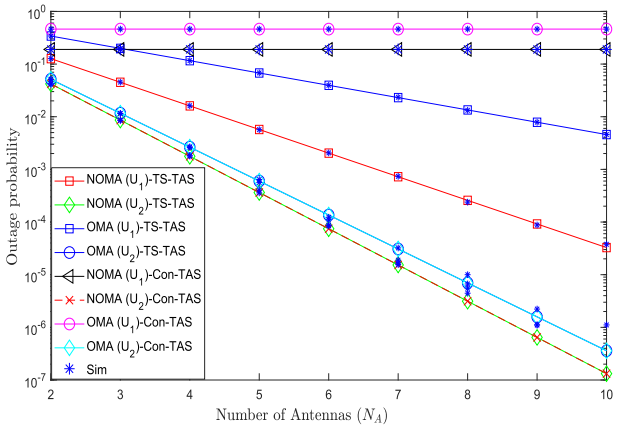

Fig. 5. Comparison between the simulation and theoretical results, and outage probabilities analysis of the considered TAS schemes as a function the number of transmit antennas N_{A}, when ρ is set to 15 dB.

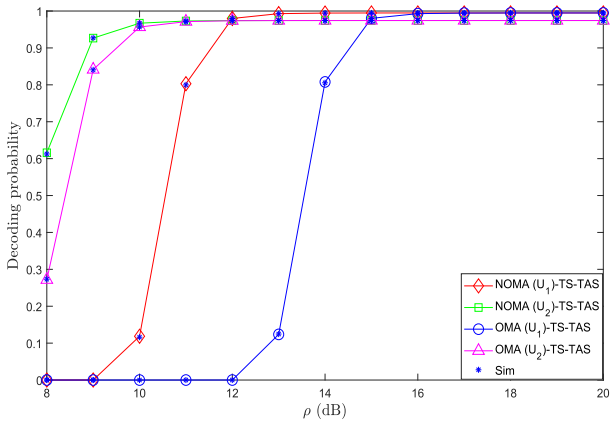

Fig. 6. Simulation results and the performance comparison between NOMA and OMA-based RLC implementation as a function of transmit power to noise ratio ρ, when the other parameters are set as: $q = 4$, $K_1 = K_2 = 100$, $M_1 = M_2 = 103$, $\hat{w} = 10$, $L = 5$ and $N_{\mathrm{A}} = 10$.

coefficients i.e., $a_1 = 0.2$ and $a_2 = 0.8$. In addition, we set $\tau_u = 1$. Unless otherwise stated, we consider uncoded BPSK for group $\mathcal{U}_1$ and convolutional coded BPSK for $\mathcal{U}_2$, which are characterized by the SNR thresholds $\hat{\gamma}_1 = 5.782$ dB and $\hat{\gamma}_2 = -0.983$ dB, respectively, given in [41] as discussed in Section II.

Fig. 5 plots the outage probabilities of the conventional TAS (referred as Con-TAS) versus the proposed two-stage TAS (referred as TS-TAS) protocol. In addition it also illustrates the relationship between the number of transmit antennas and the outage performance of the two protocols. The simulation results confirm the accuracy of analytical expressions. It can be observed that Con-TAS protocol shows optimal outage performance for U_2. However, it provides worse and constant outage probability at U_1. This is because the Con-TAS protocol only selects the best antenna for U_2, which acts as a random antenna for U_1 as discussed in Section IV. Therefore by the Con-TAS protocol, U_1 cannot benefit from multiple transmit antennas. On the other hand, TS-TAS outperforms Con-TAS protocol by reducing the outage probability at U_1 while keeping the optimal performance at U_2. Therefore, TS-TAS protocol can be adopted to meet the different QoS requirements. Furthermore, NOMA based TAS protocols show always superior performance than OMA-based TAS protocols. In addition, when the number of antennas N_{A} is increased, the gap between the outage probabilities achieved by NOMA and OMA-based protocols becomes larger.

Fig. 6 demonstrates the agreement between the simulation and analytical results, which confirms the accuracy of our derived expressions. It also exhibits the effect of the transmitted power on the decoding performance at both U_1 and U_2 for a simulation setting with $K_1 = K_2 = 100$, $L = 5$, $\hat{w} = 10$ and $N_{\mathrm{A}} = 10$. The decoding probability at U_2 is greater than the probability at U_1, because high power is allocated to U_2 in order to satisfy its quality of service requirement. In addition, coded BPSK is employed for U_2 which is less sensitive to channel errors than un-coded BPSK employed for U_1. This performance gap is even larger in OMA-based RLC implementation, because each user is allocated half of

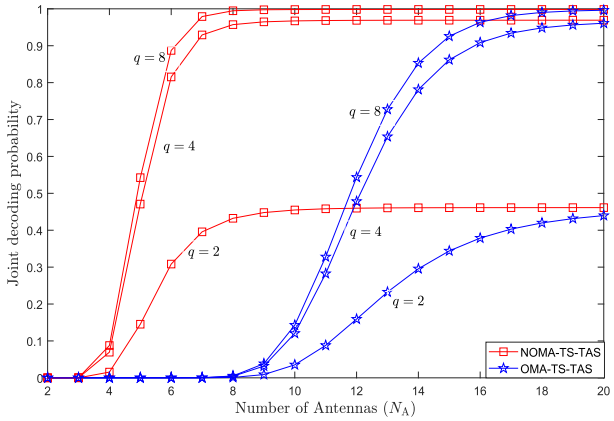

Fig. 7. Effect of finite field size q on the joint decoding probability versus the number of transmit antennas N_{A}, when $K_1 = K_2 = 100$, $L = 5$, $\hat{w} = 10$, $M_1 = M_2 = 103$ and $\rho = 13$ dB.

the total bandwidth. Thus, as expected, NOMA combined with RLC outperforms OMA-based RLC.

Fig. 7 exhibits the effect of finite field q on the joint decoding probability as a function of the number of transmit antennas, when $\rho = 13$ dB, and constant number of frames M_i have been transmitted. Note that, since the joint decoding probability is the product of two probabilities, therefore its distribution is more affected by the distribution of the minimum of the two probabilities. It can be observed that a remarkable performance gain is achieved when the field size q increases from $q = 2$ to $q = 4$. However, the performance gain is comparatively smaller when q further increases from $q = 4$ to $q = 8$. This is because the certainty of linear independence between coded packets increases with the field size and approaches the highest possible degree even for relatively small values of q. Furthermore, it can be seen that when $q = 2$ and $q = 4$, the decoding probabilities converge to a constant value because of the limited number of transmissions $M_1 = M_2 = 103$. The decoding probabilities can reach 1 if the number of transmissions increases. The figure also depicts the gain in decoding probability of NOMA-based

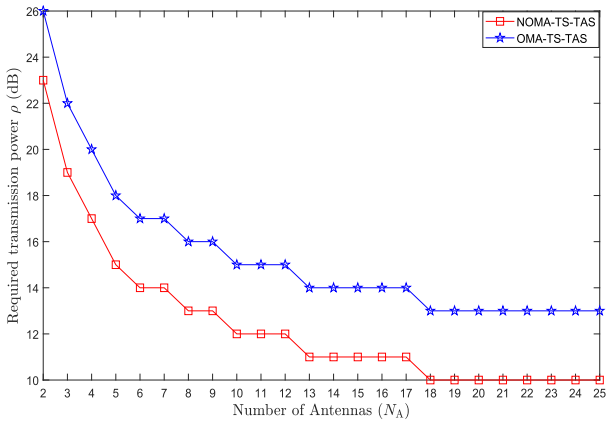

Fig. 8. Comparison between the two schemes in terms of the required transmission power versus the number of transmit antennas, when $q = 4$, $K_1 = K_2 = 100$, $M_1 = M_2 = 103$, $\hat{w} = 10$ and $L = 4$.

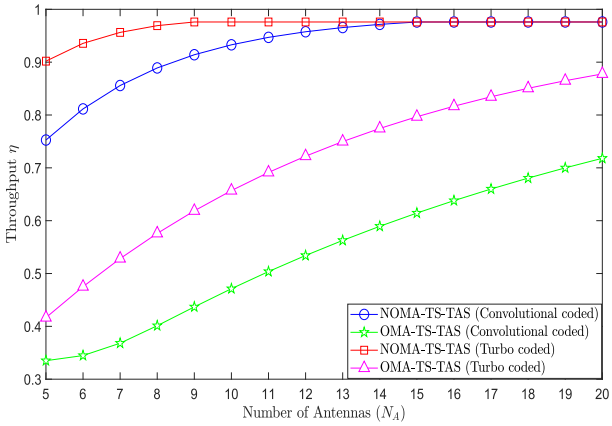

Fig. 9. Effect of coding schemes on the network throughput against the number of transmit antennas N_A. The remaining parameters of the network have been set as follows: $K_1 = K_2 = 40$, $L = 3$, $\hat{w} = 10$ and $\rho = 10$ dB.

implementation. For example, for the decoding probability to converge to the maximum possible value, the OMA-based RLC scheme requires more than twice the number of antennas in comparison to NOMA-based implementation. Alternatively, OMA requires higher transmission power to achieve the same performance gain as of NOMA-based RLC.

Fig. 8 illustrates the relationship between the number of transmit antennas and the required transmission power for the successful delivery of data files by their respective users U_1 and U_2. The figure clearly demonstrates the reduced transmission requirements offered by NOMA-based RLC as opposed to OMA with RLC. It can be observed that, for a fixed number of transmit antennas, the OMA-based implementation requires more transmission power as compared to NOMA. Alternatively as depicted in Fig. 7, OMA requires more transmit antennas in order to achieve the same performance as NOMA with RLC. At low values of ρ, it is interesting to note that OMA requires more than twice the number of antennas needed for NOMA with RLC. However at high ρ values, this difference reduces to a small number. For example, when $\rho = 13$ dB, NOMA-based RLC requires 8 or more antennas while OMA combined with RLC needs at least 19 antennas. On the other hand when ρ increases to 19 dB, only two extra antennas are required by OMA to achieve the same decoding performance as of NOMA-based implementation.

Fig. 9 presents the network throughput as a function of the number of transmit antennas and demonstrates the effect of different coding schemes employed for U_2, when the transmission power is set to a low value such that $\rho = 10$ dB. The SNR threshold for turbo coded-BPSK is set to $\hat{\gamma}_2 = -4.401$ dB, as given in [41]. As expected, turbo coded-BPSK provides a higher throughput than convolutional coded-BPSK. The performance gap between NOMA-based RLC and OMA-based RLC is evident. We can also observe that a change in the coding scheme from turbo coding to convolutional coding will cause a more notable throughput degradation in the case of OMA-based RLC than in NOMA-based RLC. Convolutional coding will make both schemes more sensitive to frame errors but throughput degradation is accentuated in OMA due to the spectral loss of 1/2, which dominates system performance.

VI. Conclusions

This paper proposed a framework which combines RLC and NOMA-based TAS schemes to support multicast groups of users with different QoS requirements and processing capabilities. Simulation results confirmed the validity of the theoretical expressions, which can be used to determine the design parameters of both RLC and NOMA schemes so that a desired QoS can be achieved. We have studied the impact of TAS on the performance of both NOMA-based RLC and conventional OMA-based RLC. We noted that, compared to the conventional TAS scheme, the two-stage TAS criterion can efficiently exploit the benefits of multiple antennas and thus can significantly improve the overall network performance. Moreover, superior performance in terms of reduced transmission power and network throughput can always be achieved by employing NOMA-based RLC than its counterpart OMA-based RLC. We also noted that network reliability can be improved further by increasing the field size over which random linear coding is performed as well as by using stronger error correction schemes at the physical layer.

Appendix

For completeness, known definitions and relationships that are invoked in the proof of the equivalence between (6) and (7) are first presented and the steps for obtaining (6) from (7) are then detailed.

The *Gaussian binomial coefficient* is the q-analog of the binomial coefficient and is defined as [60, p. 125]

$$\begin{bmatrix} N \\ K \end{bmatrix}_q = \begin{cases} \prod_{i=0}^{K-1} \frac{(q^N - q^i)}{(q^K - q^i)}, & \text{for } K \leq N \\ 0, & \text{for } K > N \end{cases} \tag{32}$$

where N, K are non-negative integers and q is a prime power. Using (3), we can rewrite (32) as

$$\begin{bmatrix} N \\ K \end{bmatrix}_q = \frac{f(N,K)}{f(K,K)}$$
$$\Leftrightarrow f(N,K) = \begin{bmatrix} N \\ K \end{bmatrix}_q f(K,K). \tag{33}$$

Combining (4) and (33) gives

$$f_r(N,K) = \left(\frac{f(N,r)}{f(r,r)}\right) f(K,r) = \begin{bmatrix} N \\ r \end{bmatrix}_q f(K,r). \tag{34}$$

Two identities that involve Gaussian binomial coefficients and will be used in the Appendix are:

$$\begin{bmatrix} N \\ r \end{bmatrix}_q = \begin{bmatrix} N \\ N-r \end{bmatrix}_q \tag{35}$$
$$\begin{bmatrix} N \\ r \end{bmatrix}_q \begin{bmatrix} r \\ K \end{bmatrix}_q = \begin{bmatrix} N \\ K \end{bmatrix}_q \begin{bmatrix} N-K \\ r-K \end{bmatrix}_q. \tag{36}$$

Another relationship that will be invoked in the subsequent proof is the decomposition of $f(N, k_1+k_2)$ into a product, as follows:

$$\begin{aligned} f(N,k_1+k_2) &= \prod_{i=0}^{k_1-1}\left(q^N - q^i\right)\prod_{i=k_1}^{k_1+k_2-1}\left(q^N-q^i\right) \\ &= f(N,k_1)\prod_{j=0}^{k_2-1}\left(q^N - q^{j+k_1}\right) \\ &= f(N,k_1)q^{k_1k_2}\prod_{j=0}^{k_2-1}\left(q^{N-k_1}-q^j\right) \\ &= f(N,k_1)\, f(N-k_1,k_2)\, q^{k_1k_2} \end{aligned} \tag{37}$$

where $j = i-k_1$ in the second line of (37). The last expression of interest is the *q-Vandermonde identity*, which states that

$$\begin{bmatrix} n_1+n_2 \\ K \end{bmatrix}_q = \sum_r \begin{bmatrix} n_1 \\ K-r \end{bmatrix}_q \begin{bmatrix} n_2 \\ r \end{bmatrix}_q q^{r(n_1-K+r)} \tag{38}$$

for $\max(0, K-n_1) \le r \le \min(n_2, K)$.

In order to prove that (6) can be obtained from (7), we first expand (7) into (10) and modify it as follows

$$\begin{aligned} &\sum_{r_1} f_{r_1}(n_2,w_1) f(n_1,k_1-r_1) q^{n_1r_1} f(n_2-r_1,k_2) q^{w_1k_2} \\ &\overset{(a)}{=} \sum_{r_1} \begin{bmatrix} n_2 \\ r_1 \end{bmatrix}_q f(w_1,r_1) f(n_1,k_1-r_1) \begin{bmatrix} n_2-r_1 \\ k_2 \end{bmatrix}_q \\ &\quad \cdot f(k_2,k_2) q^{n_1r_1+w_1k_2} \\ &\overset{(b)}{=} \sum_{r_1} \begin{bmatrix} n_2 \\ k_2 \end{bmatrix}_q \begin{bmatrix} n_2-k_2 \\ r_1 \end{bmatrix}_q f(w_1,r_1) f(n_1,k_1-r_1) \\ &\quad \cdot f(k_2,k_2) q^{n_1r_1+w_1k_2} \\ &\overset{(c)}{=} f(n_2,k_2) \sum_{r_1} \begin{bmatrix} n_2-k_2 \\ r_1 \end{bmatrix}_q f(w_1,r_1) f(n_1,k_1-r_1) \\ &\quad \cdot q^{n_1r_1+w_1k_2} \end{aligned} \tag{39}$$

where step (a) invokes (33) and (34) to expand $f(n_2-r_1,k_2)$ and $f_{r_1}(n_2,w_1)$, respectively; step (b) uses the identities (35) and (36) to rewrite the product of the two Gaussian binomial coefficients, and step (c) applies (33) to obtain $f(n_2,k_2)$. The following additional steps can further expand the product $f(w_1,r_1)f(n_1,k_1-r_1)$ that appears in (39):

$$\begin{aligned} &f(w_1,r_1)f(n_1,k_1-r_1) \\ &\overset{(d)}{=} f(w_1,r_1)f(n_1-k_1+w_1,\, w_1-r_1) \\ &\quad \cdot f(n_1,k_1-w_1) q^{(k_1-w_1)(w_1-r_1)} \\ &\overset{(e)}{=} f(w_1,r_1) \begin{bmatrix} n_1-k_1+w_1 \\ w_1-r_1 \end{bmatrix}_q f(w_1-r_1,w_1-r_1) \\ &\quad \cdot f(n_1,k_1-w_1) q^{(k_1-w_1)(w_1-r_1)} \\ &\overset{(f)}{=} f(w_1,w_1) \begin{bmatrix} n_1-k_1+w_1 \\ w_1-r_1 \end{bmatrix}_q f(n_1,k_1-w_1) \\ &\quad \cdot q^{-r_1(w_1-r_1)+(k_1-w_1)(w_1-r_1)} \end{aligned} \tag{40}$$

where step (d) first adds and subtracts w_1 in order to rewrite $f(n_1,k_1-r_1)$ as $f(n_1,[k_1-w_1]+[w_1-r_1])$ and then applies (37); step (e) uses (33) to expand $f(n_1-k_1+w_1,\, w_1-r_1)$, and step (f) first multiplies and divides all terms by $f(w_1,w_1)$ and then observes that

$$\frac{f(w_1,r_1)\, f(w_1-r_1,w_1-r_1)}{f(w_1,w_1)} = \frac{\prod_{i=0}^{w_1-r_1-1}(q^{w_1-r_1}-q^i)}{\prod_{i=r_1}^{w_1}(q^{w_1}-q^i)} = q^{-r_1(w_1-r_1)}.$$

Substitution of (40) into (39) gives

$$\begin{aligned} &\sum_{r_1} f_{r_1}(n_2,w_1) f(n_1,k_1-r_1) q^{n_1r_1} f(n_2-r_1,k_2) q^{w_1k_2} \\ &= f(n_1,k_1-w_1)\, f(n_2,k_2)\, q^{w_1(k_1+k_2-w_1)} \\ &\quad \cdot f(w_1,w_1) \sum_{r_1} \begin{bmatrix} n_1-k_1+w_1 \\ w_1-r_1 \end{bmatrix}_q \begin{bmatrix} n_2-k_2 \\ r_1 \end{bmatrix}_q q^{r_1(n_1-k_1+r_1)} \\ &\overset{(g)}{=} f(n_1,k_1-w_1)\, f(n_2,k_2)\, q^{w_1(k_1+k_2-w_1)} \\ &\quad \cdot f(w_1,w_1) \begin{bmatrix} n_1+n_2-k_1-k_2+w_1 \\ w_1 \end{bmatrix}_q \\ &\overset{(h)}{=} f(n_1,k_1-w_1) f(n_2,k_2) f(N-K+w_1,w_1) q^{w_1(K-w_1)} \end{aligned} \tag{41}$$

where step (g) invokes (38) and step (h) uses (33) to combine $f(w_1,w_1)$ with the Gaussian binomial coefficient and obtain $f(N-K-w_1,w_1)$, given that $k_1+k_2=K$ and $n_1+n_2=N$. We note that

$$\begin{aligned} f(N-K+w_1,w_1) q^{w_1(K-w_1)} &= \prod_{i=0}^{w_1-1} q^{K-w_1}(q^{N-K+w_1}-q^i) \\ &= \prod_{j=K-w_1}^{K-1}(q^N-q^j) \end{aligned} \tag{42}$$

for $j = i+K-w_1$. If we substitute (42) into (41), we obtain expression (6).

REFERENCES

[1] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 1617–1655, 3rd Quart., 2016.

[2] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. IEEE Veh. Technol. Conf. (VTC-Spring)*, Dresden, Germany, Jun. 2013, pp. 1–5.

[3] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. Bhargava, "A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2181–2195, Oct. 2017.

[4] Z. Ding *et al.*, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

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

[6] M. Al-Imari, P. Xiao, M. A. Imran, and R. Tafazolli, "Uplink non-orthogonal multiple access for 5G wireless networks," in *Proc. Int. Symp. Wireless Commun. Syst. (ISWCS)*, Barcelona, Spain, 2014, pp. 781–785.

[7] A. Goldsmith, *Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[8] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Pers. Commun.*, vol. 6, no. 3, pp. 311–335, Mar. 1998.

[9] İ. E. Telatar, "Capacity of multi-antenna Gaussian channels," *Eur. Trans. Telecommun.*, vol. 10, no. 6, pp. 585–595, 1999.

[10] S. Sanayei and A. Nosratinia, "Antenna selection in MIMO systems," *IEEE Commun. Mag.*, vol. 42, no. 10, pp. 68–73, Oct. 2004.

[11] N. B. Mehta, S. Kashyap, and A. F. Molisch, "Antenna selection in LTE: From motivation to specification," *IEEE Commun. Mag.*, vol. 50, no. 10, pp. 144–150, Oct. 2012.

[12] Z. Chen, J. Yuan, and B. Vucetic, "Analysis of transmit antenna selection/maximal-ratio combining in Rayleigh fading channels," *IEEE Trans. Veh. Technol.*, vol. 54, no. 4, pp. 1312–1321, Jul. 2005.

[13] Y. Zhang, J. Ge, and E. Serpedin, "Performance analysis of nonorthogonal multiple access for downlink networks with antenna selection over Nakagami-m fading channels," *IEEE Trans. Veh. Technol.*, vol. 66, no. 11, pp. 10590–10594, Nov. 2017.

[14] Y. Yu, H. Chen, Y. Li, Z. Ding, L. Song, and B. Vucetic, "Antenna selection for MIMO nonorthogonal multiple access systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3158–3171, Apr. 2018.

[15] X. Liu and X. Wang, "Efficient antenna selection and user scheduling in 5G massive MIMO-NOMA system," in *Proc. IEEE Veh. Technol. Conf. (VTC Spring)*, Nanjing, China, May 2016, pp. 1–5.

[16] Y. Yu, H. Chen, Y. Li, Z. Ding, and L. Zhuo, "Antenna selection in MIMO cognitive radio-inspired NOMA systems," *IEEE Commun. Lett.*, vol. 21, no. 12, pp. 2658–2661, Dec. 2017.

[17] T. Ho, M. Médard, J. Shi, M. Effros, and D. Karger, "On randomized network coding," in *Proc. 41st Allerton Conf. Commun., Control, Comput.*, Monticello, IL, USA, Oct. 2003, pp. 1–10.

[18] T. Ho *et al.*, "A random linear network coding approach to multicast," *IEEE Trans. Inf. Theory*, vol. 52, no. 10, pp. 4413–4430, Oct. 2006.

[19] D. Szabo, A. Gulyas, F. H. P. Fitzek, and D. E. Lucani, "Towards the tactile Internet: Decreasing communication latency with network coding and software defined networking," in *Proc. 21st Eur. Wireless Conf.*, Budapest, Hungary, May 2015, pp. 1–6.

[20] J. Zhang, P. Fan, and K. B. Letaief, "Network coding for efficient multicast routing in wireless ad-hoc networks," *IEEE Trans. Commun.*, vol. 56, no. 4, pp. 598–607, Apr. 2008.

[21] R. Bassoli, H. Marques, J. Rodriguez, K. W. Shum, and R. Tafazolli, "Network coding theory: A survey," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 5, pp. 1950–1978, 4th Quart., 2013.

[22] Y. Yang, C. Zhong, Y. Sun, and J. Yang, "Energy efficient reliable multi-path routing using network coding for sensor network," *Int. J. Comput. Sci. Netw. Secur.*, vol. 8, no. 12, pp. 329–338, Dec. 2008.

[23] A. G. Dimakis, P. B. Godfrey, Y. Wu, M. J. Wainwright, and K. Ramchandran, "Network coding for distributed storage systems," *IEEE Trans. Inf. Theory*, vol. 56, no. 9, pp. 4539–4551, Sep. 2010.

[24] A. S. Khan and I. Chatzigeorgiou, "Opportunistic relaying and random linear network coding for secure and reliable communication," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 223–234, Jan. 2018.

[25] Y. Lin, B. Liang, and B. Li, "Priority random linear codes in distributed storage systems," *IEEE Trans. Parallel Distrib. Syst.*, vol. 20, no. 11, pp. 1653–1667, Nov. 2009.

[26] D. Vukobratović and V. Stanković, "Unequal error protection random linear coding strategies for erasure channels," *IEEE Trans. Commun.*, vol. 60, no. 5, pp. 1243–1252, May 2012.

[27] N. Thomos, E. Kurdoglu, P. Frossard, and M. van der Schaar, "Adaptive prioritized random linear coding and scheduling for layered data delivery from multiple servers," *IEEE Trans. Multimedia*, vol. 17, no. 6, pp. 893–906, Jun. 2015.

[28] M. Esmaeilzadeh, P. Sadeghi, and N. Aboutorab, "Random linear network coding for wireless layered video broadcast: General design methods for adaptive feedback-free transmission," *IEEE Trans. Commun.*, vol. 65, no. 2, pp. 790–805, Feb. 2017. [Online]. Available: http://arxiv.org/abs/1411.1841

[29] A. Tassi, I. Chatzigeorgiou, and D. Vukobratović, "Resource-allocation frameworks for network-coded layered multimedia multicast services," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 2, pp. 141–155, Feb. 2015.

[30] C. W. Sørensen, D. E. Lucani, F. H. P. Fitzek, and M. Médard, "On-the-fly overlapping of sparse generations: A tunable sparse network coding perspective," in *Proc. IEEE Veh. Technol. Conf. (VTC Fall)*, Vancouver, BC, USA, Sep. 2014, pp. 1–5.

[31] P. Maymounkov, N. J. A. Harvey, and D. S. Lun, "Methods for efficient network coding," in *Proc. 44th Allerton Conf. Commun., Control Comput.*, Monticello, IL, USA, Sep. 2006, pp. 482–491.

[32] Y. Li, E. Soljanin, and P. Spasojević, "Effects of the generation size and overlap on throughput and complexity in randomized linear network coding," *IEEE Trans. Inf. Theory*, vol. 57, no. 2, pp. 1111–1123, Feb. 2011.

[33] A. Heidarzadeh and A. H. Banihashemi, "Overlapped chunked network coding," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Cairo, Egypt, Jan. 2010, pp. 1–5.

[34] M. C. O. Bogino, P. Cataldi, M. Grangetto, E. Magli, and G. Olmo, "Sliding-window digital fountain codes for streaming of multimedia contents," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, New Orleans, LA, USA, May 2007, pp. 3467–3470.

[35] Y. Lin, B. Liang, and B. Li, "SlideOR: Online opportunistic network coding in wireless mesh networks," in *Proc. IEEE INFOCOM*, San Diego, CA, USA, Mar. 2010, pp. 1–5.

[36] D. J. C. MacKay, "Fountain codes," *IEE Proc.-Commun.*, vol. 152, no. 6, pp. 1062–1068, Dec. 2005.

[37] S. Feizi, D. E. Lucani, and M. Médard, "Tunable sparse network coding," in *Proc. Int. Zurich Seminar Commun. (IZS)*, Zürich, Switzerland, Feb. 2012, pp. 107–110.

[38] S. Park and D.-H. Cho, "Random linear network coding based on non-orthogonal multiple access in wireless networks," *IEEE Commun. Lett.*, vol. 19, no. 7, pp. 1273–1276, Jul. 2015.

[39] A. S. Khan and I. Chatzigeorgiou, "Non-orthogonal multiple access combined with random linear network coded cooperation," *IEEE Signal Process. Lett.*, vol. 24, no. 9, pp. 1298–1302, Sep. 2017.

[40] C.-Y. Tu, C.-Y. Ho, and C.-Y. Huang, "Energy-efficient algorithms and evaluations for massive access management in cellular based machine to machine communications," in *Proc. IEEE Veh. Technol. Conf.*, Sep. 2011, pp. 1–5.

[41] I. Chatzigeorgiou, I. J. Wassell, and R. Carrasco, "On the frame error rate of transmission schemes on quasi-static fading channels," in *Proc. 42nd Conf. Inf. Sci. Syst.*, Princeton, NJ, USA, Mar. 2008, pp. 577–581.

[42] Y. Xi, A. Burr, J. Wei, and D. Grace, "A general upper bound to evaluate packet error rate over quasi-static fading channels," *IEEE Trans. Wireless Commun.*, vol. 10, no. 5, pp. 1373–1377, May 2011.

[43] I. Chatzigeorgiou, I. J. Wassell, and R. Carrasco, "Threshold-based frame error rate analysis of MIMO systems over quasistatic fading channels," *Electron. Lett.*, vol. 45, no. 4, pp. 216–217, Feb. 2009.

[44] T. Liu, L. Song, Y. Li, Q. Huo, and B. Jiao, "Performance analysis of hybrid relay selection in cooperative wireless systems," *IEEE Trans. Commun.*, vol. 60, no. 3, pp. 779–788, Mar. 2012.

[45] C.-C. Chao, C.-C. Chou, and H.-Y. Wei, "Pseudo random network coding design for IEEE 802.16m enhanced multicast and broadcast service," in *Proc. IEEE Veh. Technol. Conf. (VTC-Spring)*, Taipei, Taiwan, May 2010, pp. 1–5.

[46] J. Heide, M. V. Pedersen, F. H. P. Fitzek, and M. Médard, "On code parameters and coding vector representation for practical RLNC," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kyoto, Japan, Jun. 2011, pp. 1–5.

[47] J. H. van Lint and R. M. Wilson, *A Course in Combinatorics*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2001.

[48] È. M. Gabidulin, "Theory of codes with maximum rank distance," *Problems Inf. Transmiss.*, vol. 21, no. 1, pp. 1–12, Jan. 1985.

[49] M. Gadouleau and Z. Yan, "Constant-rank codes and their connection to constant-dimension codes," *IEEE Trans. Inf. Theory*, vol. 56, no. 7, pp. 3207–3216, Jul. 2010.

[50] J. Qureshi, C. H. Foh, and J. Cai, "Primer and recent developments on fountain codes," *Recent Patents Telecommun.*, vol. 2, no. 1, pp. 2–11, Jul. 2013.
[51] S. Feizi, D. E. Lucani, C. W. Sørensen, A. Makhdoumi, and M. Médard, "Tunable sparse network coding for multicast networks," in *Proc. Int. Symp. Netw. Coding*, Aalborg, Denmark, Jun. 2014, pp. 1–6.
[52] A. Tassi, I. Chatzigeorgiou, and D. E. Lucani, "Analysis and optimization of sparse random linear network coding for reliable multicast services," *IEEE Trans. Commun.*, vol. 64, no. 1, pp. 285–299, Jan. 2016.
[53] P. Cataldi, M. Grangetto, T. Tillo, E. Magli, and G. Olmo, "Sliding-window raptor codes for efficient scalable wireless video broadcasting with unequal loss protection," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1491–1503, Jun. 2010.
[54] J. K. Sundararajan, D. Shah, M. Médard, S. Jakubczak, M. Mitzenmacher, and J. Barros, "Network coding meets TCP: Theory and implementation," *Proc. IEEE*, vol. 99, no. 3, pp. 490–512, Mar. 2011.
[55] (2013). *The NWCRG Website*. [Online]. Available: https://irtf.org/nwcrg
[56] P. A. Chou, Y. Wu, and K. Jain, "Practical network coding," in *Proc. 41st Allerton Conf. Commun., Control Comput.*, Monticello, IL, USA, Oct. 2003, pp. 40–49.
[57] A. Papoulis and S. U. Pillai, *Probability, Random Variables and Stochastic Processes*. New York, NY, USA: McGraw-Hill, 2002.
[58] I. Chatzigeorgiou and A. Tassi, "Decoding delay performance of random linear network coding for broadcast," *IEEE Trans. Veh. Technol.*, vol. 66, no. 8, pp. 7050–7060, Aug. 2017.
[59] X. Wang, W. Chen, and Z. Cao, "SPARC: Superposition-aided rateless coding in wireless relay systems," *IEEE Trans. Veh. Technol.*, vol. 60, no. 9, pp. 4427–4438, Nov. 2011.
[60] P. J. Cameron, *Combinatorics: Topics, Techniques, Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 1994.

Amjad Saeed Khan received the B.Eng. degree in computer engineering from the COMSATS Institute of Information Technology, Pakistan, in 2010, and the M.Sc. degree in digital signal processing and intelligent systems and the Ph.D. degree in communication systems from Lancaster University, U.K., in 2013 and 2018, respectively. He is currently a Research Associate in signal processing for 5G networks with the Wolfson School of Mechanical, Electrical and Manufacturing Engineering, Loughborough University, U.K. His research interests include 5G networks, network coding, secure wireless communication, digital signal processing, nonorthogonal multiple access, embedded systems design, machine learning, and blockchain technology.

Ioannis Chatzigeorgiou (M'05–SM'15) received the Dipl.Ing. degree in electrical engineering from the Democritus University of Thrace, Greece, the M.Sc. degree in satellite communication engineering from the University of Surrey, U.K., and the Ph.D. degree from the University of Cambridge, U.K. He is currently a Senior Lecturer with the School of Computing and Communications, Lancaster University, U.K. Prior to his appointment, he was with Marconi Communications and Inmarsat Ltd. He also held a Post-doctoral position at the University of Cambridge and the Norwegian University of Science and Technology supported by the Engineering and Physical Sciences Research Council and the European Research Consortium for Informatics and Mathematics, respectively. His research interests include communication theory, co-operative networks, relay-aided communications, and network coding.

Sangarapillai Lambotharan (SM'06) received the Ph.D. degree in signal processing from Imperial College London, London, in 2017. He was a Post-Doctoral Research Associate with Imperial College London until 1999. He was a Visiting Scientist with the Engineering and Theory Centre, Cornell University, USA, in 1996. From 1999 to 2002, he was with the Motorola Applied Research Group, U.K., where he investigated various projects including physical link layer modeling and performance characterization of GPRS, EGPRS, and UTRAN. He was a Lecturer with Kings College London and a Senior Lecturer with Cardiff University from 2002 to 2007. He is currently a Professor of digital communications and the Head of the Signal Processing and Networks Research Group, Wolfson School Mechanical, Electrical and Manufacturing Engineering, Loughborough University, Loughborough, U.K. He has authored over 200 journals and conference articles in his research areas. His current research interests include 5G networks, MIMO, radars, smart grids, machine learning, network security, and blockchain technology.

Gan Zheng (S'05–M'09–SM'12) received the B.Eng. and M.Eng. degrees in electronic and information engineering from Tianjin University, Tianjin, China, in 2002 and 2004, respectively, and the Ph.D. degree in electrical and electronic engineering from The University of Hong Kong in 2008. He is currently a Reader of signal processing for wireless communications with the Wolfson School of Mechanical, Electrical and Manufacturing Engineering, Loughborough University, U.K. His research interests include machine learning for communications, UAV communications, mobile edge caching, full-duplex radio, and wireless power transfer. He was a first recipient of the 2013 IEEE SIGNAL PROCESSING LETTERS Best Paper Award, and he also received the 2015 GLOBECOM Best Paper Award and the 2018 Best Paper Award from the IEEE Technical Committee on Green Communications & Computing. He currently serves as an Associate Editor for the IEEE COMMUNICATIONS LETTERS.