IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 23, 2019, accepted June 29, 2019, date of publication July 3, 2019, date of current version July 22, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2926770

Flex-NOMA: Exploiting Buffer-Aided Relay Selection for Massive Connectivity in the 5G Uplink

NIKOLAOS NOMIKOS[1], (Member, IEEE), EMMANOUEL T. MICHAILIDIS[2], PANAGIOTIS TRAKADAS[3], DEMOSTHENES VOUYIOUKAS[1], (Senior Member, IEEE), THEODORE ZAHARIADIS[3], AND IOANNIS KRIKIDIS[4], (Fellow, IEEE)

[1]Information and Communication Systems Engineering Department, University of the Aegean, 83200 Samos, Greece
[2]Electrical and Electronics Engineering Department, University of West Attica, Campus 2, 12244 Aigaleo, Greece
[3]General Department, National and Kapodistrian University of Athens, 34400 Psahna, Greece
[4]Electrical and Computer Engineering Department, University of Cyprus, 1678 Nicosia, Cyprus

Corresponding author: Nikolaos Nomikos (nnomikos@aegean.gr)

This work was supported in part by the framework of the DEFENDER and NRG5 projects, in part by the European Commission through the Horizon 2020 and 5G-PPP programmes under Grant ICT-740898 and ICT-762013, and in part by the European Regional Development Fund and the Republic of Cyprus through the Research and Innovation Foundation under Project INFRASTRUCTURES/1216/0017 (IRIDA).

ABSTRACT The success of fifth generation (5G) mobile network and its long-term evolution (LTE) relies on enabling demanding services with massive connectivity requirements, including Internet of Things (IoT) nodes, mobile devices, or unmanned aerial vehicles (UAVs). Towards this end, non-orthogonal multiple access (NOMA) schemes allow multiple users and devices to communicate using the same temporal and spectral resources. In this context, buffer-aided (BA) relay selection can significantly enhance the quality and reliability of communication, through increased diversity. So, in this paper, we employ BA relay selection in the uplink of the NOMA networks where users and devices coexist and demand access to the wireless resources. The presented relay selection policy, namely flex − NOMA, facilitates simultaneous transmissions from multiple sources to multiple relays, exploiting channel state information at the reception, and dynamic decoding ordering by the relays performing successive interference cancellation (SIC). The theoretical analysis and performance evaluation results are provided and comparisons, in terms of outage probability, average sum-rate and average delay show that flex − NOMA offers improved performance without incurring high complexity and coordination overheads.

INDEX TERMS NOMA, relay selection, buffer-aided relaying, H2H, IoT, diversity.

I. INTRODUCTION

Fifth generation (5G) networks introduce new services, based on Internet of Things (IoT) devices and unmanned aerial vehicles (UAVs), that current networks are struggling to support due to limited spectrum resources. A promising technique, allowing massive connectivity is non-orthogonal multiple access (NOMA) [1]. Contrary to orthogonal multiple access (OMA), NOMA allows multiple devices to share resources, either through power domain (PD) or code domain (CD) multiplexing [2]. PD NOMA is based on superposition coding at the transmitters and successive interference cancellation (SIC) at the receivers. The survey in [3] presented an overview of PD NOMA and solutions for its efficient implementation, discussing the impact of user channel asymmetry, fairness and cooperation. In this context, various cooperative paradigms have been proposed, either through user cooperation or with dedicated relays, significantly improving the overall coverage [4], [5]. So, novel cooperation algorithms for massive connectivity of coexisting Human-to-Human (H2H) and IoT devices is a pre-requisite for enabling 5G services. In this area, several open issues have been identified, motivating the research in the current paper, including the different Quality-of-Service (QoS) of H2H and IoT services and the excessive overhead, calling for low-complexity implementation [6].

The associate editor coordinating the review of this manuscript and approving it for publication was Yuanwei Liu.

A. RELATED WORK

In order to satisfy 5G capacity demands, the downlink and uplink should be carefully designed. In greater detail, power allocation, hardware capabilities for SIC, user clustering and coordination in dense environments are pivotal when developing NOMA solutions [7], [8]. In recent years, there have been significant contributions on NOMA algorithms, with the majority of works focusing on the downlink (see e.g., [9], [10] and references therein).

In the uplink of NOMA networks, a major issue is how to determine the user decoding order, achieving efficient NOMA operation. The authors in [11], [12] studied dynamic decoding order reception, based on the instantaneous received power. So, firstly, the signal of the strongest user was decoded, while finally, the weakest user was decoded without any interference. Performance evaluation showed reduced outages compared to fixed ordering. The authors in [13] examined dynamic user clustering and power allocation for both downlink and uplink NOMA by formulating a sum-throughput maximization problem. Thus, a low-complexity user clustering solution was developed, surpassing OMA for various number of users. For massive IoT, the work in [14] presented random NOMA, where each device randomly selected a channel and the base station performed SIC. Also, the optimal resource allocation and the system stability condition were given. Then, a random access scheme was given for coexisting NOMA and OMA machine-to-machine (M2M) communication, determining the number of devices for each one. The authors in [15] studied multi-cell uplink NOMA with multiple relays, achieving a diversity gain per symbol equal to two, without channel state information at the transmitter (CSIT).

An important technique for enhanced cooperation is opportunistic relay selection (ORS). In OMA, ORS has received numerous contributions (see e.g., [16], [17] and references therein), outlining its central role for increasing diversity and throughput, while reducing outages. In NOMA, various ORS works, focused on the downlink. The paper in [18] proposed a two-stage ORS algorithm, showing that outage probability was minimized, outperforming conventional max − min selection and ORS OMA. When CSIT is partially available, i.e., CSIT is known only in the source-relay ($\{S\rightarrow R\}$) or relay-destination ($\{R\rightarrow D\}$) links, the authors in [19] presented partial relay selection (PRS). Performance evaluation illustrated that for low and medium SNR, PRS increased the ergodic capacity, as more relays were added to the network, contrary to cases with high SNR. Next, dual relay selection was developed in [20], relying on distributed space-time coding. Lower and upper outage bounds were derived and numerical results showed that compared to single relay selection, outages were reduced. Moreover, two ORS algorithms were proposed in [21], for relays with full-duplex (FD) and half-duplex (HD) capabilities. Analytical outage expressions were derived, outlining the improved performance of FD ORS for low SNR, while HD ORS was superior for high SNR.

Building upon ORS, buffer-aided (BA) relaying enabled flexible relay selection. For OMA, several works studied BA ORS (see e.g., [22], [23] and references therein), while recently, BA ORS has been investigated for NOMA. For a single relay network, the authors in [24] proposed BA NOMA relaying, with and without CSIT in the $\{R\rightarrow D\}$ link. Also, link selection was presented, achieving a diversity order of two for buffers larger than two packets. In a similar setting, the authors in [25] presented an adaptive algorithm where the relay switched between NOMA and OMA in the $\{R\rightarrow D\}$ link. Results showed improved sum-throughput and higher complexity, compared to both OMA and NOMA without buffering. Then, the study in [26] maximized a long-term network utility through the joint optimization of transmit mode, admission control and power allocation, considering buffer stability and power constraints. For multi-relay networks, BA ORS was presented in [27] for reducing the average delay and maintaining the diversity. Also, power allocation for NOMA in the $\{R\rightarrow D\}$ link was performed, providing increased robustness against outages. Next, NOMA with multi-relay reception was investigated in [28], extending previous BA ORS algorithms for OMA [29], [30]. Two BA ORS algorithms were given, using buffer state information (BSI), activating the relay with the maximum number of packets for transmission. Results showed that the hybrid NOMA/OMA algorithm, increased the average sum-rate and reduced the average delay, compared to OMA and NOMA. Finally, the paper in [31] examined a multi-relay two user network and proposed a BA ORS policy for IoT services. Prioritization-based selection with NOMA/OMA was presented, achieving a diversity equal to three times the number of relays.

B. MOTIVATION AND CONTRIBUTIONS

Although BA ORS for NOMA has received adequate contributions, most works have focused on the downlink. Nonetheless, the uplink of multi-relay NOMA networks provides an important research field with several challenges that must be addressed. Firstly, resource constrained devices, in terms of hardware, require channel access and distributed power control might be infeasible on such devices. Thus, low-complexity implementation is required to accommodate the needs of heterogeneous devices. Secondly, power control in uplink NOMA is not trivial, as allocating power to specific users to attain channel asymmetry at one relay might degrade the SIC performance at another relay. As a result, current relay selection solutions should be adapted to uplink NOMA, improving the chances for SIC at one or more relays. Finally, as low-latency is a key 5G performance indicator, buffers increase the required time for a packet to reach the destination. So, BA ORS should activate relays, according to BSI and prioritize $\{R\rightarrow D\}$ transmissions for reducing packet delays. Here, BA ORS for uplink NOMA is presented, exploiting multi-relay reception and dynamic decoding ordering at the $\{S\rightarrow R\}$ link, as well as BSI for relay selection at the $\{R\rightarrow D\}$ link. In addition, low-complexity implementation is

promoted, as SIC is performed at the relays without requiring any CSIT. More specifically, the following contributions are provided:

1) A BA ORS policy, namely flex − NOMA for multi-relay uplink NOMA networks, is presented. Flex − NOMA allows simultaneous transmissions from multiple sources and relay selection, based on BSI and $\{R \rightarrow D\}$ link prioritization for reduced packet delays. Also, increased flexibility is provided by flex − NOMA, as packets from a subset of the sources are stored at the relays, while rate adaptation is employed in the $\{R \rightarrow D\}$ transmission.

2) Outage analysis for flex − NOMA is presented, where the link outage probabilities and the network outage probabilities are given. The analysis considers the effect of dynamic decoding ordering at the relays, and the impact of buffers in allowing either the $\{S \rightarrow R\}$ or the $\{R \rightarrow D\}$ link to be activated.

3) Performance evaluation in terms of outage probability, average sum-rate and average delay, as well as comparisons with an equivalent BA ORS OMA policy are given. From the results, it is shown that flex − NOMA outperforms OMA, without demanding increased complexity, being capable of supporting the 5G uplink requirements.

C. OUTLINE

The paper is organized as follows. In Section II, the system model is given, while Section III presents flex − NOMA. Next, Section IV, includes an outage analysis for flex − NOMA and then, Section V includes comparisons between the proposed, as well as relevant selection policies. Finally, Section VI discusses the conclusions and future directions.

II. SYSTEM MODEL

A. NETWORK MODEL

A relay-assisted network is considered, consisting of N sources, S_n $(1 \leq n \leq N)$, one destination, D, and a cluster $\mathcal{C}$ of K HD decode-and-forward (DF) relays, $R_k \in \mathcal{C}$ $(1 \leq k \leq K)$. Due to severe fading, the direct links between the sources and the destination do not exist and it is assumed that communication can only be established via the relays. Each relay R_k is equipped with a buffer of size L, where L denotes the maximum number of data elements that can be stored from the source's transmissions. The number of packets in the buffer of relay R_k is denoted by Q_k. Each buffer is allocated equally to each source i.e., the same amount of data elements of $S_1, S_2, \ldots, S_N$ can be stored at the relays. This assumption results in the formation of sub-buffers, denoted as Q_{k,S_n} and the respective sub-buffer sizes, assumed to be equal at each relay are denoted by L_{S_n}. The multi-source multi-relay network is depicted in Fig. 1.

The source nodes are assumed to be saturated (i.e., they always have data to transmit) and the required information rate, r_{S_n}, for successful reception at each relay, is fixed and may differ, depending on the application; for example, if S_1 is a cellular user and S_2 is an IoT device, the rate requirements differ and, hence, $r_{S_1} \neq r_{S_2}$. Equivalently, a transmission from a transmitter i to its corresponding receiver j is successful if the SNR Γ_{ij} at the receiver is greater than or equal to a threshold γ_{ij}, called the *capture ratio*. More specifically, γ_{ij} is defined as $\gamma_{ij} = 2^{r_i - 1}$, where the value of r_i depends on the modulation and coding characteristics of the application. At each time-slot, the sources or one of the relays attempt to transmit a packet, using a fixed power level P_i, $i \in \{S_1, \ldots, S_N, R_1, \ldots, R_K\}$.

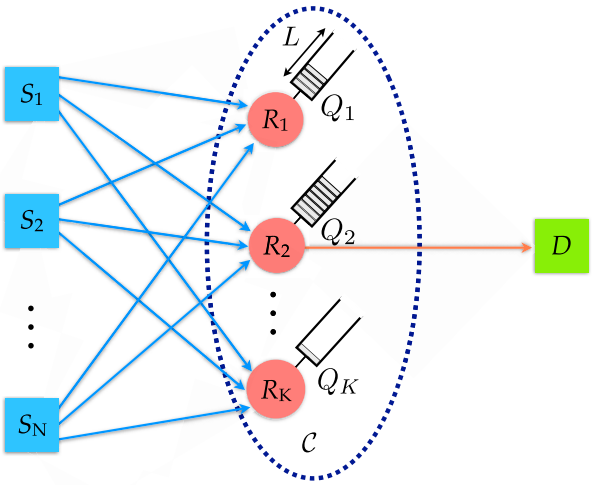

FIGURE 1. Sources $S_1, S_2, \ldots, S_N$ use NOMA to communicate in the uplink with the base station D via a cluster of relays $R_k \in \mathcal{C}$, $k \in \{1, 2, \ldots, K\}$.

The retransmission process is based on an acknowledgement/negative-acknowledgement (ACK/NACK) mechanism, in which short-length error-free packets are broadcasted by the receivers over a separate narrow-band channel. In addition, as more than one relays might receive the same packet, it is necessary to notify them on which packet(s) was received by the destination. Thus, the short-length ACK packets include the packet ID information, in order for the relays to drop the respective packet(s) from their queue and avoid duplicate transmissions at another time-slot.

B. CHANNEL MODEL

Time is divided into "slots" of one packet duration. At any arbitrary time-slot t, the quality of the wireless channels is degraded by additive white Gaussian noise (AWGN) and frequency non-selective Rayleigh block fading, according to a complex Gaussian distribution with zero mean and variance σ_{ij}^2 for the $\{i \rightarrow j\}$ link. The complex channel coefficient for the $\{i \rightarrow j\}$ link is denoted by h_{ij}, and the channel gain, $g_{ij} \triangleq |h_{ij}|^2$, is assumed to be exponentially non-identically distributed i.e., $g_{ij} \sim \text{Exp}(\lambda_{ij})$, $\lambda_{ij} > 0$, as is the case of an asymmetric topology. The variance of thermal noise at a receiver l is denoted by σ_l^2, $l \in \{R_1, R_2, \ldots, R_K, D\}$ and it is assumed

to be distributed as AWGN, with zero mean and variance equal to σ_l^2.

1) TRANSMISSION IN THE $\{S \rightarrow R\}$ LINK

When transmission in the $\{S \rightarrow R\}$ link occurs, the information symbols of the N sources, i.e., $x_1, \ldots, x_N$ with $\mathbb{E}[|x_n^2|] = 1$, $n \in \{1, 2, \ldots, N\}$ are transmitted concurrently, through NOMA.

Then, relay R_k will receive signal y_k, containing the symbols of the N sources, given by

$$y_k = \sum_{n=1}^{N} h_{S_n R_k} \sqrt{P_{S_n}} x_n + \eta_{R_k}; \tag{1}$$

where η_{R_k} denotes the AWGN at relay R_k.

Regarding SIC operation, the signal of S_n can be successfully decoded at R_k if

$$\Gamma_{S_n R_k}(P_{S_n}) \triangleq \frac{g_{S_n R_k} P_{S_n}}{\sum_{i=n+1}^{N} g_{S_i R_k} P_{S_i} + \sigma_{R_k}^2} \geq 2^{r_{S_n}} - 1; \tag{2}$$

assuming that the previous $N - n$ signals have been successfully decoded at R_k, being subtracted from y_k prior to decoding the signal of S_n.

For ordering the sources' signals, the dynamic SIC receiver of [12] is employed. This receiver relies on CSI at the relays to perform ordering of the signals of the sources, based on the instantaneous received signal power. By denoting as Φ the set of all possible decoding orders, R_k determines the permutation ϕ_k, $\phi_k \in \Phi$ with which, the sources' signals are ordered. After an arbitrary broadcasting phase, each R_k sequentially decodes the signals by ordering the sources as $\phi_{k,1}, \phi_{k,2}, \ldots, \phi_{k,N}$, according to their respective channel gains $g_{\phi_{k,1} R_k} \geq g_{\phi_{k,2} R_k} \geq \cdots \geq g_{\phi_{k,N} R_k}$, as equal transmit power level is assumed at every source. So, the strongest signal is decoded first, by considering the other $N-1$ signals as interference. Then, SIC subtracts the decoded signal and moves on to the second strongest signal. This process is repeated until the weakest signal of the source with index $\phi_{k,N}$ is decoded interference-free. Also, the set of feasible $\{S \rightarrow R\}$ links is denoted by $\mathcal{F}_{SR}^n$, having a cardinality of F_{SR}^n, where the respective relays have successfully decoded packets of n sources after the dynamic SIC receiver operation, i.e., its members are the links fulfilling (2).

Remark 1: It must be noted that although a BA relay network is assumed here, SIC operation does not depend on whether or not there is available space to store the data of sources whose signals are decoded at the relays' buffers. More specifically, if a packet is successfully decoded during SIC but the buffer space allocated to the respective source is full, the packet is subtracted by SIC and discarded by the respective relay.

2) TRANSMISSION IN THE $\{R \rightarrow D\}$ LINK

In practical scenarios, each source might transmit at a different rate r_{S_n}, as diverse applications are supported, especially in networks with coexisting users and devices. Then, in order to avoid overflow or starvation of the buffers, the selected relay R_k will forward a combined packet to the destination, transmitting at a maximum rate $r_{\max} = \sum_{n=1}^{N} r_{S_n}$, i.e., a rate equal to the sum of the rate requirements of the N sources, adjusting and extending the NOMA network topology with one source and two users in [8]. Therefore, if N packets are combined, the SNR at the destination should fulfill

$$\Gamma_{R_k D}(P_{R_k}) \triangleq \frac{g_{R_k D} P_{R_k}}{\sigma_D^2} \geq 2^{r_{\max}} - 1. \tag{3}$$

On the contrary, the combined packet, containing the data of N sources will not be successfully transmitted to the destination, if $\gamma_{R_k D} < 2^{r_{\max}} - 1$. As a result, the probability of outage for the simultaneous transmission of the data of N sources by R_k, in the respective $\{R \rightarrow D\}$, is expressed as follows

$$p_{\text{out}\{R \rightarrow D\}} \triangleq \Pr\left[g_{R_k D} < \frac{(2^{r_{\max}} - 1)\sigma_D^2}{P_{R_k}} \right]. \tag{4}$$

As will be discussed at Section III-B, additional flexibility to the wireless transmission is enabled, since when the proposed BA ORS policy is adopted, the number of packets being stored and transmitted is not strictly equal to N. More specifically, in terms of channel capacity, the worst case is equivalent to having a wireless channel supporting only the transmission of the packet of the source maintaining a service with the minimum rate requirement. In this case, the probability in (4) does not depend on $r_{\max}$, but on $r_{\min}$, where $r_{\min} = \min\{r_1, r_2, \ldots, r_N\}$ corresponds to the minimum rate requirement, in order to avoid a complete outage in the $\{R \rightarrow D\}$ link.

Since in the considered topology, CSIT in the $\{R \rightarrow D\}$ links is assumed to be available, the selected relay employs rate adaptation, choosing a transmission rate, aiming to transmit packets from as many sources as possible. By $\mathcal{F}_{RD}^n$, the set of feasible $\{R \rightarrow D\}$ links is denoted, having a cardinality of F_{RD}^n, where the respective relays can transmit packets from n sources, that currently reside in their buffers. On the contrary, if a $\{R \rightarrow D\}$ link is feasible, i.e., it can support the transmission of a packet with rate r_{S_i}, $i \in \{1, 2, \ldots, N\}$ but the relay's buffer is empty, that link is assumed to be in outage.

III. BUFFER-AIDED RELAY SELECTION FOR UPLINK NOMA

A. RELEVANT POLICIES

Below, relevant BA ORS policies are presented, aiming to offer a clear perspective on the operation of two-hop BA multi-relay networks. The following three BA ORS policies provide the necessary means for the development of flex − NOMA for the uplink of NOMA networks.

1) MAX-LINK: ADAPTIVE LINK AND RELAY SELECTION

Early BA relay selection policies operated using a two-slot cooperative protocol with fixed scheduling for the source and relay transmission. However, this limitation was relaxed in [23], allowing each slot to be dynamically allocated to

either $\{S\rightarrow R\}$ or $\{R\rightarrow D\}$ transmission, based on the instantaneous link quality and the status of the relays' buffers. More specifically, max − link relay selection leverages the flexibility of buffering at the relays. Thus, in order to activate the best link, max − link compares the channel gains of the available links, allocating each time-slot to the strongest link.

1. If a $\{S\rightarrow R\}$ link prevails, the source transmits and one relay is chosen for reception. A $\{S\rightarrow R\}$ link is considered to be available for selection when the respective relay's buffer is not full and hence, it can receive the source's data.
2. On the other hand, if a $\{R\rightarrow D\}$ link is the strongest one, the corresponding relay is activated for transmission. A $\{R\rightarrow D\}$ link is considered to be available as long as the respective relay's buffer is not empty and thus, there exists source data that can be forwarded to the destination.

The max − link relay selection policy is expressed as follows:

$$R^* = \arg\max_{R_k\in\mathcal{C}} \left\{ \bigcup_{R_k\in\mathcal{C}:\Psi(Q_k)\neq L} \{g_{SR_k}\}, \bigcup_{R_k\in\mathcal{C}:\Psi(Q_k)\neq 0} \{g_{R_kD}\} \right\}, \tag{5}$$

where R^* denotes the selected relay for transmission or reception and the function $0 \leq \Psi(Q_k) \leq L$, provides the number of data elements that are stored in buffer Q_k.

2) LoCo-LINK: MULTI-RELAY RECEPTION, BSI EXPLOITATION AND $\{R\rightarrow D\}$ PRIORITIZATION

The LoCo − Link selection algorithm was proposed in [30] and aimed at improving the performance of BA relaying in terms of average end-to-end delay, without introducing high implementation complexity. The algorithm operates as follows:

1. Contrary to max − link, where the selection of the best link was performed among the $2K$ available ones, LoCo − Link prioritizes the $\{R\rightarrow D\}$ link, by selecting in each time-slot, the $\{R\rightarrow D\}$ link belonging in $\mathcal{F}_{RD}$, i.e., the set of feasible $\{R\rightarrow D\}$ links, having at the same time the maximum queue length. If more than one relays have the same maximum queue length, one of them is randomly activated.
2. If no $\{R\rightarrow D\}$ link is available due to severe fading or because all buffers are empty, the source transmits its packets to all the relays in the first hop. So, there might exist more than one relays forming the set $\mathcal{F}_{SR}$, containing the feasible $\{S\rightarrow R\}$ links, being able to receive and store the source's packet. As a result, in the next time-slot, the possibility of activating an $\{R\rightarrow D\}$ link is increased compared to schemes with $\{R\rightarrow D\}$ prioritization and single relay reception in the $\{S\rightarrow R\}$ link, such as the delay-aware algorithm of [32] where in addition, BSI is not exploited during relay selection.

LoCo − Link provides several gains to the two-hop BA network, reducing the average end-to-end delay without degrading the diversity of the network. In addition, when a distributed implementation is adopted, the issue of relay selection with outdated CSI can be fully mitigated.

3) BA–NOMA/OMA: FLEXIBLE MULTIPLE ACCESS

In NOMA downlink networks with one source, multiple BA relays and two destinations, hybrid BA–NOMA/OMA [28] adjusts LoCo − Link's operation for NOMA and combines both multiple-access schemes. The operation of BA–NOMA/OMA is described below.

1. In order to achieve NOMA, the BA–NOMA/OMA policy allows the source to transmit a packet with rate $r_1 + r_2$ towards the K relays, containing the data of both destinations. Then, the relays perform NOMA in the power-domain, and through power allocation, based on the instantaneous CSI of the $\{R\rightarrow D\}$ links, they transmit a superposition-coded packet to the two destinations. Nonetheless, the NOMA transmission might not be successful, i.e., due to low available transmit SNR, negligible channel asymmetry between the destinations or high rate requirements. In these cases, the network can avoid a complete outage, if at least, one destination D_j is able to receive a packet from the relays belonging in $\mathcal{F}_{RD_j}, j \in \{1, 2\}$, i.e., the set of links that can support an OMA transmission towards D_j.
2. Under OMA operation, and in order to achieve user fairness, at odd time-slots, first, D_1 is examined on whether or not it can receive its packet from $R_k \in \mathcal{F}_{RD_1}$, otherwise D_2 is examined on whether or not it can receive its packet from $R_k \in \mathcal{F}_{RD_2}$. Likewise, at even time-slots the order with which each destination is investigated, is reversed. When there exist relays with equal queue lengths in the $\{R\rightarrow D\}$ links, one of them is randomly selected.

BA–NOMA/OMA constitutes an efficient selection algorithm, combining the merits of both multiple-access schemes, and due to its increased flexibility it is able to maintain the average sum-rate of the two-hop network.

B. FLEX-NOMA RELAY SELECTION POLICY

Here, the BA relay selection policy for the uplink of NOMA multi-source multi-relay networks is presented. More specifically, the details of relay selection with additional flexibility (flex − NOMA) in storing and forwarding packets, relying on NOMA with multiple simultaneous transmissions in the uplink by N sources are given. The proposed policy is based on low complexity implementation, being especially useful in networks where resource constrained devices coexist with users and so, the required intelligence for SIC and BSI exploitation relay selection is transferred at the relay nodes.

1) $\{R\rightarrow D\}$ LINK

Targeting to guarantee low end-to-end average delay, $\{R\rightarrow D\}$ transmissions are prioritized and relay selection exploits the availability of BSI. More specifically, as already discussed at Section II, in each buffer, the space is equally allocated to each source, thus being divided in N sub-buffers. So, as

sub-buffers are formed, the relay having the maximum sub-buffer size from the set of relays, storing packets from the maximum number of sources is chosen to transmit to the destination. This relay forwards packets coming from n different sources to the destination, where $n \leq N$. Adopting such a policy, where BSI is considered in the selection, allows the selected relay to avoid overflowing the sub-buffer space of one or more sources, thus preserving the diversity of the network. In addition, in the $\{R \rightarrow D\}$ link, CSIT is assumed to be available, as it is required at each relay to determine the amount of data that can be successfully transmitted to the destination. So, although a relay might fulfill the BSI criterion of flex − NOMA, its $\{R \rightarrow D\}$ link might not be able to support the data rate for successful delivery of the combined packet to the destination.

2) $\{S \rightarrow R\}$ LINK

However it might occur that no relay is able to transmit with at least $r_{\min}$ towards the destination, either due to empty buffers or severe fading in the $\{R \rightarrow D\}$ links, simultaneous $\{S \rightarrow R\}$ transmissions by the N sources take place. Instead of relying on fixed user ordering, each relay exploits CSI at the reception to determine the decoding order, thus maximizing the probability for successful SIC. More specifically, flex − NOMA allows BA relays to store packets, even when SIC is not able to decode the packets of all N sources. Each relay stores the successfully decoded packets from different sources, considering that the respective sub-buffers are not full. As already discussed, each buffer is equally divided to store the packets of each source. Thus, as flex − NOMA allows packets from a subset of the sources to be stored and transmitted, it is possible that the number of packets, residing in each sub-buffer will be different.

3) FLEX-NOMA OPERATION

It is well-known that the performance of NOMA networks relies on pairing together asymmetric users, in the sense of either channel conditions or rate requirements. So, in the considered topology, the availability of multiple relays, provides increased degrees of freedom (DoF) for successful SIC at one or more relays. More importantly, ORS can leverage these DoF with low complexity, employing broadcasting by the sources without power control or highly complex pairing algorithms, thus avoiding CSIT in the $\{S \rightarrow R\}$ links where the majority of network nodes exist. Thus, flex − NOMA promotes low complexity implementation, since it requires only BSI and CSIT at the relay to operate. More specifically, in the $\{S \rightarrow R\}$ links, dynamic user ordering at the relays requires only local CSI, in order to form the set $\mathcal{F}_{SR}^{n}$, while CSIT is used in the $\{R \rightarrow D\}$ links to determine the links belonging in set $\mathcal{F}_{RD}^{n}$. Moreover, as a multi-relay topology is considered, power control of the transmit power employed by the sources to enhance channel asymmetries might not be practical, as improving the chances of SIC at one relay might degrade the performance of SIC at another relay. As a result, in flex − NOMA, each source transmits with a fixed power level, assumed to be equal for all the sources. Next, Algorithm 1 describes the operation of flex − NOMA at an arbitrary time-slot:

Algorithm 1 flex − NOMA Relay Selection

input $\mathcal{F}_{RD}^{n}, n \in \{1, 2, \ldots, N\}$
if $\mathcal{F}_{RD}^{n} = \emptyset, \forall n$ **then**
  The N sources broadcast their packets to the K relays.
  $Q_j \leftarrow Q_j + m, \quad \forall j \in \mathcal{F}_{SR}^{m}, \quad m \in \{1, 2, \ldots, N\}$
else
  $n' = \arg\max_n \mathcal{F}_{RD}^{n}$
  $i' = \arg\max_{i \in \mathcal{F}_{RD}^{n'}} Q_{i,S_l}$
  if more than one relays have the same maximum sub-buffer length **then**
    i^* is chosen randomly from the set of relays in i'.
  else
    $i^* = i'$
  end if
  $Q_i^* \leftarrow Q_i^* - n$
end if
Output Link $\{R_{i^*} \rightarrow D\}$ is activated to transmit with rate $r = \sum_{i=1}^{n} r_i$ to the destination or the set of links in $\mathcal{F}_{SR}^{m}$ receive m packets from N sources through NOMA broadcasting, where $m \leq N$.

Remark 2: It should be highlighted that flex − NOMA *can be implemented in either centralized or distributed fashion. In the centralized approach, D is responsible to acquire the CSI and BSI of all K relays and perform relay selection at each time-slot. On the contrary, in the distributed approach, relay selection relies on the use of timers at the relays. More specifically, the relays that are not in outage to D, set the values of their timers according to the number of packets from different sources that can be transmitted to D and the maximum sub-buffer length. Similar schemes have been proposed in [16], [30], and [33] for the distributed implementation of relay selection policies in the downlink of multi-relay networks with a single source and a single destination.*

IV. OUTAGE ANALYSIS

A. PER LINK OUTAGE PROBABILITY

1) $\{S \rightarrow R\}$ LINK OUTAGE PROBABILITY

The successful reception of the signals, being simultaneously transmitted by the N sources, depends on the channel asymmetries among the $\{S \rightarrow R\}$ links and the decoding order determined by each relay R_k. Here, the framework of [11], [12] is adopted where each receiver determines dynamically the decoding order of the sources' signals, based on their instantaneous received power. Assuming that R_k has adopted a decoding order $\phi_k \in \Phi$, the rate of source $\phi_{k,n}$ will be given by

$$r_{\phi_{k,n}R_k} \triangleq \log_2\left(1+\frac{g_{\phi_{k,n}R_k}P_{\phi_{k,n}}}{\sum_{i=n+1}^{N} g_{\phi_{k,i}R_k}P_{\phi_{k,i}} + \sigma_{R_k}^2}\right). \quad (6)$$

Moreover, the probability of successful transmission when a rate $r_{\phi_{k,n}}$ is required is formulated as

$$\Pr(E_{\phi_{k,n}}) = \Pr(r_{\phi_{k,n}R_k} \geq r_{\phi_{k,n}}). \tag{7}$$

When $r_{\phi_{k,n}}$ cannot be supported, the outage probability is

$$\begin{aligned} p_{\text{out}\{\phi_{k,n}\to R_k\}} &\triangleq 1 - \prod_{\phi_{k,i}=1}^{\phi_{k,n}} \Pr(E_{\phi_{k,i}}) \\ &= 1 - \prod_{\phi_{k,i}=1}^{\phi_{k,n}} \Pr(r_{\phi_{k,i}R_k} \geq r_{\phi_{k,i}}). \end{aligned} \tag{8}$$

So, the outage probability for the transmission of S_n for all possible decoding orders at R_k is

$$p_{\text{out}\{S_n\to R_k\}} = \sum_{\phi_k\in\Phi, S_n=\phi_{k,n}} \Pr(\phi_k) \times p_{\text{out}\{\phi_{k,n}\to R_k\}}; \tag{9}$$

where $\Pr(\phi_k)$ denotes the probability of the decoding order ϕ_k. Equivalently, the respective success probability is denoted by $p_{\{S_n\to R\}} = 1 - p_{\text{out}\{S_n\to R_k\}}$.

Then, assuming a topology with three sources performing simultaneous transmissions, at R_k, the probability of the decoding order ϕ_k is equal to [11, (7)]

$$\Pr(\phi_k) \triangleq \frac{\lambda_{\phi_{k,2}}\lambda_{\phi_{k,3}}}{(\lambda_{\phi_{k,1}} + \lambda_{\phi_{k,2}})(\lambda_{\phi_{k,1}} + \lambda_{\phi_{k,2}} + \lambda_{\phi_{k,3}})}. \tag{10}$$

Furthermore, closed form expressions for the probability of successful decoding of the signal of each user given a specific decoding order $\phi_{k,n}$ and various rate requirements, have been derived in [11]. Focusing on the case where each source demands a rate $r_{\phi_{k,n}} \geq 1$bps/Hz, $n \in \{1, 2, 3\}$, corresponding to $\gamma_{\phi_{k,n}R_k} \geq 1$, and after simple algebraic manipulations and adjustment of the expressions given in [11] to the case of multi-relay reception, $\Pr(E_{\phi_{k,1}})$, $\Pr(E_{\phi_{k,2}})$ and $\Pr(E_{\phi_{k,3}})$ can expressed as

$$\Pr(E_{\phi_{k,1}}) = \frac{\alpha\beta}{\delta\epsilon} e^{-\lambda_{\phi_{k,1}}\sigma^2_{R_k}\gamma_{\phi_{k,1}R_k}}; \tag{11}$$

$$\Pr(E_{\phi_{k,2}}) = \frac{\alpha\beta}{\zeta} e^{-(\lambda_{\phi_{k,1}}+\lambda_{\phi_{k,2}})\sigma^2_{R_k}\gamma_{\phi_{k,2}R_k}}; \tag{12}$$

$$\Pr(E_{\phi_{k,3}}) = e^{-(\lambda_{\phi_{k,1}}+\lambda_{\phi_{k,2}}+\lambda_{\phi_{k,3}})\sigma^2_{R_k}\gamma_{\phi_{k,3}R_k}}; \tag{13}$$

where $\alpha = \lambda_{\phi_{k,1}} + \lambda_{\phi_{k,2}}$, $\beta = \lambda_{\phi_{k,1}} + \lambda_{\phi_{k,2}} + \lambda_{\phi_{k,3}}$, $\delta = \lambda_{\phi_{k,1}}\gamma_{\phi_{k,1}R_k} + \lambda_{\phi_{k,2}}$, $\epsilon = 2\lambda_{\phi_{k,1}}\gamma_{\phi_{k,1}R_k} + \lambda_{\phi_{k,2}} + \lambda_{\phi_{k,3}}$ and $\zeta = \lambda_{\phi_{k,1}}\gamma_{\phi_{k,2}R_k} + \lambda_{\phi_{k,2}}\gamma_{\phi_{k,2}R_k} + \lambda_{\phi_{k,3}}$. Inserting (11)-(13) in (8), (9), the outage probability of the transmission in the $\{S_n \to R_k\}$ link can be derived.

For the considered scenario, assuming that the wireless links are independent and non-identically distributed (i.n.i.d.), the outage probability of S_n over all K relays and decoding orders is given by

$$p_{\text{out}\{S_n\to R\}} = \sum_{k=1}^{K} p_{\text{out}\{S_n\to R_k\}}. \tag{14}$$

Remark 3: It should be noted that for an N-source single-hop NOMA network with a single receiver, performing dynamic user ordering, closed-form outage expressions have been derived in [12]. Readers are referred to that work for further details on the analysis of such a topology.

2) {$R\to D$} LINK OUTAGE PROBABILITY

In the $\{R\to D\}$ link, since the relays have full CSIT knowledge, they can perform rate adaptation. As a result, depending on link quality, the selected relay will transmit at a rate level between $r_{\min} = \min\{r_1, r_2, \ldots, r_N\}$ and $r_{\max} = \sum_{n=1}^{N} r_n$. Here, an outage event is denoted when the selected relay cannot guarantee a transmission at rate $r_{\min}$. So, a $\{R\to D\}$ link is in outage if $\gamma_{R_kD} < 2^{r_{\min}} - 1$, and the probability of outage is given by

$$\begin{aligned} p_{\text{out}\{R_{k,S_{r_{\min}}}\to D\}} &\triangleq \Pr\left[g_{R_kD} < \frac{(2^{r_{\min}} - 1)\sigma_D^2}{P_{R_k}}\right] \\ &= 1 - e^{-\lambda_{R_kD}\frac{(2^{r_{\min}}-1)\sigma_D^2}{P_{R_k}}}; \end{aligned} \tag{15}$$

since $g_{R_kD} \sim \text{Exp}(\lambda_{R_kD})$, $\lambda_{R_kD} > 0$.

Regarding the transmissions at rates above $r_{\min}$, their respective outage probabilities depend on the desired rate level, consisting of the rate requirements of the various sources. As long as R_k stores packets from multiple sources, the outage probability can be calculated by replacing $r_{\min}$ in (15) with the sum of the respective rate requirements. Correspondingly, the success probability when the rate of a source i can be supported by R_k is denoted by $p_{\{R_{k,i}\to D\}} = 1 - p_{\text{out}\{R_{k,i}\to D\}}$.

B. NETWORK OUTAGE PROBABILITY

In flex − NOMA, if successful decoding of all the N packets is not feasible, packets that were decoded by each relay are stored at the respective buffer. Then, in the $\{R\to D\}$ link, depending on the channel conditions, the relay that can transmit the maximum number of packets and has the largest queue size is selected. Thus, as it is evident from the description of flex − NOMA, the number of packets that enter the network and the number of packets that leave the network are not necessarily equal at each time-slot. As a result, in flex − NOMA, these data cannot be considered as one packet, as at a relay, the data might be separated and only the data that were successfully decoded are stored, while later, depending on $\{R\to D\}$ channel gains, different data might be transmitted to the destination. Thus, here, the theoretical analysis differs, although it can be cast as a discrete time Markov chain (DTMC) in which, each state represents a possible state of the buffers.

1) STATES OF THE DTMC

As the DTMC states show all possible buffer states, a specific state is denoted by $S_r \in \mathcal{S}$, where $\mathcal{S}$ is the set of all available states, $r \in \mathbb{N}$, $1 \leq r \leq |\mathcal{S}|$, and $|\mathcal{S}|$ is the cardinality of all the possible buffer state combinations.

2) CONSTRUCTION OF THE STATE TRANSITION MATRIX OF THE DTMC

Let $(X_t)_{t\geq 0}$ denote the discrete-time Markov random process of system evolution. Also, let $\mathbf{A} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ denote the state transition matrix of the DTMC, where entry $\mathbf{A}_{i,j} = \Pr(S_j \rightarrow S_i) \triangleq \Pr(X_{t+1} = S_i | X_t = S_j)$ corresponds to the transition probability of moving from state S_j at a time-slot t to state S_i at a time-slot $(t+1)$. At each time-slot, the buffer state and, thus, the system state, is modified as follows: (a) the number of elements of one or more buffers can be decreased by n, $n \leq N$, if a relay node is activated to transmit n packets and the transmission is successful, while other relays having the same or part of the transmitted packets, drop them from their buffers; (b) the number of elements of the buffer of an arbitrary relay R_k can be increased by $m_k, m_k \leq N$, if the sources are activated for broadcasting and after SIC, m_k packets are decoded; (c) the buffer states do not change when all the $\{S\rightarrow R\}$ and $\{R\rightarrow D\}$ links are in outage, not being able to support the transmission of a packet of a single source.

3) PROPERTIES OF THE DTMC

The DTMC represents finite-size buffers and, so, it is stationary, irreducible and aperiodic (SIA) [23], i.e., a steady state (also known as the distribution of the DTMC) $\boldsymbol{\lambda}$ exists such that $\mathbf{A}\boldsymbol{\lambda} = \boldsymbol{\lambda}$ and $\mathbf{b}^T\boldsymbol{\lambda} = 1$, where $\mathbf{b} = \left[1\ 1 \dots 1\right]^T$. For the column stochastic matrix $\mathbf{A}$ of the MC modeling the system states, $\boldsymbol{\lambda}$ can be found in [23, Lemma 1] as $\boldsymbol{\lambda} = (\mathbf{A} - \mathbf{I} + \mathbf{B})^{-1}\mathbf{b}$, where $\mathbf{I}$ is the identity matrix and $\mathbf{B}$ is a square matrix with $\mathbf{B}_{ij} = 1$ for all i, j, each having appropriate dimensions. The analytical expressions for the outage probability, as well as other metrics of interest, i.e., average throughput and average packet delay, for any finite-length BA relay selection policy, based on constructing the DTMC are given in several papers; see, e.g., [30, Equations (7), (8) and (11)].

4) ILLUSTRATIVE EXAMPLE ($N = 3, K = 1, L = 6$)

In order to better show the operation of flex − NOMA, an illustrative example is presented, focusing on the transition probabilities of state S_{111}, i.e., the state where one packet from each source resides in the relay's buffer. Also, in this example, S_1 and S_2 require rates $r_{\min} = r_{S_1} = r_{S_2} = 1$ bps/Hz, while S_3 demands $r_{S_3} = 3$ bps/Hz. Moreover, the success probabilities of transmitting in the $\{S\rightarrow R\}$ and $\{R\rightarrow D\}$ links with the required rates are clearly depicted in the transition probabilities of Fig. 2. Furthermore, in this topology, there are $K(L/N+1)^N = 27$ possible states, however, based on flex − NOMA, as presented in Section III-B.3, S_{111} can be connected to 14 states, as depicted in Fig. 2. Also, for the transitions due to $\{R\rightarrow D\}$ transmissions, a relay i (here $i = 1$, as a single-relay case is assumed), transmitting packets of sources j, k, l is denoted as $R_{i,jkl}$. By expanding the state transition diagram to include all the possible states, the state transition matrix $\mathbf{A}$ and the network outage probability can be easily calculated.

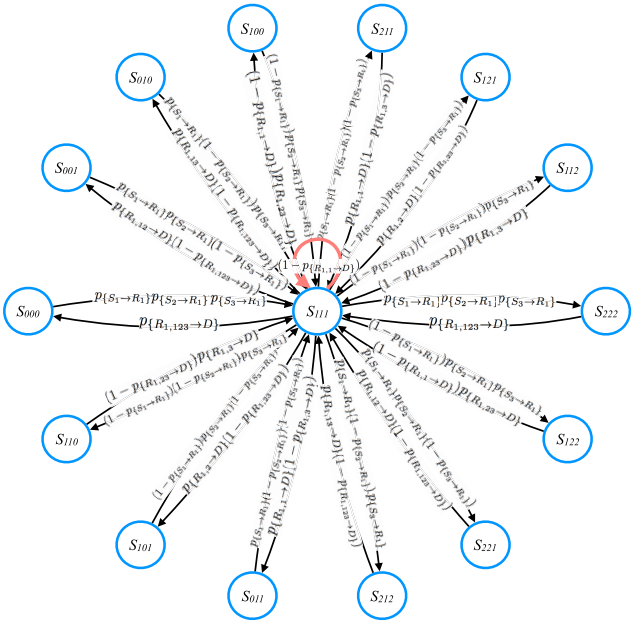

FIGURE 2. Transition diagram of state S_{111}.

V. PERFORMANCE EVALUATION

In this section, simulations are conducted to evaluate the performance of flex − NOMA. More specifically, a network with three sources is considered, where a cellular user, denoted as S_1 transmits with rate $r_{H2H} = r_{S_1} = 3$ bps/Hz and two IoT devices, denoted as S_2, S_3 transmit with rates $r_{IoT} = r_{S_2} = r_{S_3} = 1$ bps/Hz. The $\{S\rightarrow R\}$ channels are assumed to be i.n.i.d. in order to examine the impact of dynamic decoding at the relays, while the $\{R\rightarrow D\}$ channels are assumed to be independent and identically distributed (i.i.d.) An equivalent OMA policy is assumed where at each time-slot, data transmission of a specific source is scheduled, either in the $\{S\rightarrow R\}$ or in the $\{R\rightarrow D\}$ links. For fair comparison, the required rate for successful transmission by each source is set to three times the rate of NOMA, where at each time-slot, all three sources transmit their data at the same time. Finally, in order to provide further insights on the performance of flex − NOMA, various cases of K, L and r_{IoT} are assumed.

A. OUTAGE PROBABILITY

Fig. 3 provides the outage comparison for flex − NOMA and the equivalent OMA scheme for the considered multi-relay topology with three sources. Overall, it can be observed that there is a floor in the outage performance of flex − NOMA, as the effect of interference remains, even for asymptotically high SNR. This behaviour is in line with other works, focusing on NOMA networks with dynamic user ordering at the receiver [11], [12]. Starting from the two IoT devices, it can be seen that they exhibit an almost identical performance and the small difference can be attributed to the i.n.i.d. channels. Regarding the performance of NOMA and

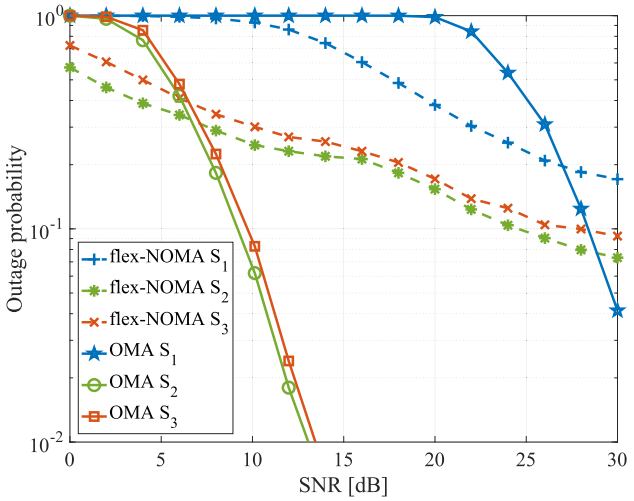

FIGURE 3. Outage probability for $K = 4$ and $L_{S_1} = L_{S_2} = L_{S_3} = 10$.

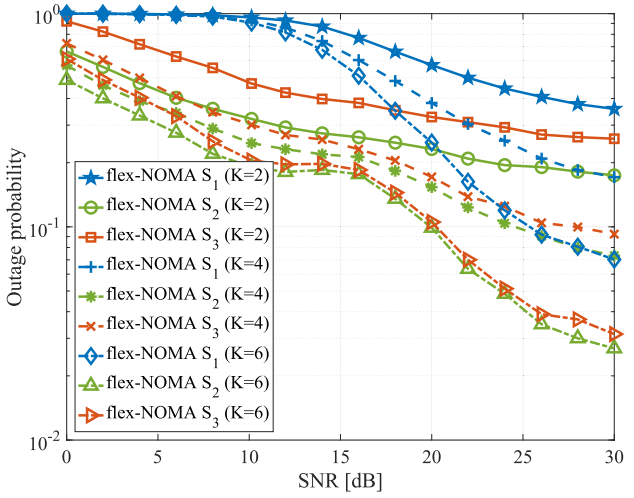

FIGURE 4. Outage probability for varying K and $L_{S_1} = L_{S_2} = L_{S_3} = 10$.

OMA, for low SNR, NOMA outperforms OMA, since for each source, the latter requires three times the rate of NOMA for transmitting three packets, once every three time-slots. Nonetheless, after 7 dB, OMA exploits the diversity of the network and due to its interference-free reception, it is able to surpass NOMA. A similar behaviour is seen for S_1 with the only difference being the increased SNR range in which OMA cannot guarantee the desired transmission rate.

The results in Fig. 4 show the outage probability performance for different values of K for flex − NOMA. Overall, it can be seen that independently of the number of available relays, an outage floor exists for flex − NOMA. Also, each case sees a floor in the medium SNR regime. This trend can be justified by considering each hop independently. Starting from the $\{R \rightarrow D\}$ links, as SNR conditions improve, outages reduce, since interference does not exist. However, after 10 dB, outages stemming from $\{R \rightarrow D\}$ transmissions tend to zero. On the contrary, in the $\{S \rightarrow R\}$ link interference remains, causing outages. Then, as SNR increases, the $\{S \rightarrow R\}$ outage probability reduces, leading the network to overcome the first floor. Next, focusing on the behaviour of each user, firstly, the cellular user, S_1 exhibits the highest outage probability, as its rate requirement is significantly higher than that of the IoT devices, S_2 and S_3. Nonetheless, as K increases, the performance gap between the three sources reduces, thus highlighting the importance of adopting multi-relay reception, increasing the chances for successful SIC at more relays. Moreover, the gains of using multiple relays in such topologies become clear, since the outage probability reduces each time the relays increase. Regarding the IoT devices, the observed differences in their outage performance derive from the i.n.i.d. $\{S \rightarrow R\}$ channels.

Then, Fig. 5 presents the outage probability performance of the cellular user S_1 when different rates are demanded by the IoT sources, S_2, S_3. The motivation behind this comparison is to outline the impact of varying the rate requirement of the IoT devices on the outage performance of the cellular user. Such a case could be mapped to a multi-user network where the relays are allocated to serve different groups of users, forming NOMA pairs. So, here, the cellular user can be paired with IoT devices requiring rates equal to $r_{IoT} = 0.5$ bps/Hz or $r_{IoT} = 1$ bps/Hz. In greater detail, when each IoT device requires a rate $r_{IoT} = 0.5$ bps/Hz, being served simultaneously with the cellular user, the probability for successful reception of S_1's signal significantly increases. On the contrary, when r_{IoT} increases, more outages occur for the cellular user. Again, the beneficial effect of adding more relays to the network is clearly observed, as the outage performance improves each time. In conclusion, these results clearly show the importance of selecting which devices to serve on the same radio-resources, in networks where coexisting cellular users and IoT devices concurrently require channel access through NOMA, necessitating the development of advanced pairing algorithms.

B. AVERAGE SUM-RATE

In the next comparison, the average end-to-end sum-rate performance is examined. More specifically, this comparison depicts the average end-to-end sum rate at the destination, where its value for flex − NOMA is determined through (2), (3) where in (3), r_{max} can be substituted with any rate combination, using the rate values required by the sources, as long as the wireless channel can support the rate combination. Fig. 6 includes results for various values of K for flex − NOMA and OMA. Starting from $K = 2$, one may observe that the fixed user scheduling of OMA results in reduced sum-rate in the low and medium SNR regime, as the cellular user is often in outage. More specifically, in OMA, S_1 must transmit with three times its required rate, once every three time-slots, in order to achieve the transmission of the same number of packets, as flex − NOMA. Nonetheless, when SNR increases, S_1 performance improves and so does

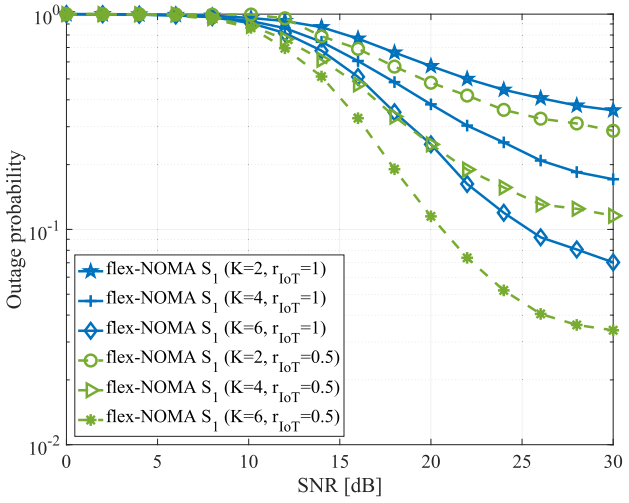

FIGURE 5. Outage probability of S_1 for various K, $L_{S_1} = L_{S_2} = L_{S_3} = 10$ and different S_2, S_3 rate requirements.

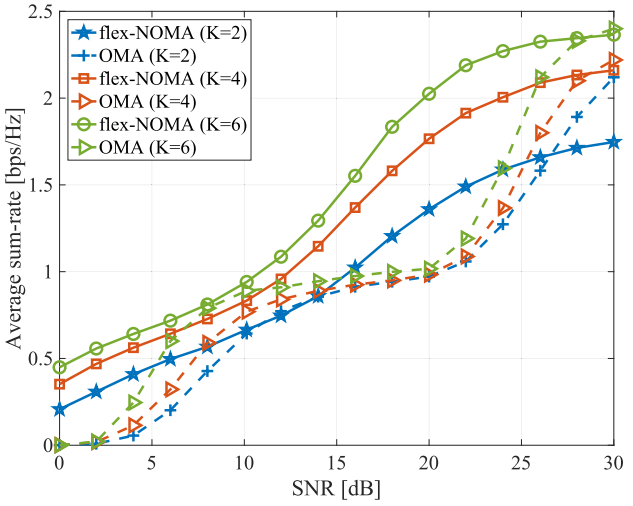

FIGURE 6. Average sum-rate for varying K and $L_{S_1} = L_{S_2} = L_{S_3} = 10$.

the average sum-rate of OMA. Moreover, for asymptotically high SNR, OMA surpasses flex − NOMA, as the latter cannot fully mitigate the interference, as it has been already seen in the outage results. Nonetheless, when more relays are available, flex − NOMA exploits the increased chances of successful SIC at the relays and closes the performance gap with OMA.

C. AVERAGE DELAY

Next, Fig. 7 focuses on the average delay of flex − NOMA and OMA for $K = 4$ and $L_{S_1} = L_{S_2} = L_{S_3} = 10$. One can observe that independently of the multiple access scheme, the delay performance of all three sources converges to one time-slot in the high SNR regime. This can be justified by the $\{R \rightarrow D\}$ proritization that is targeted in flex − NOMA and its equivalent OMA scheme, in order to

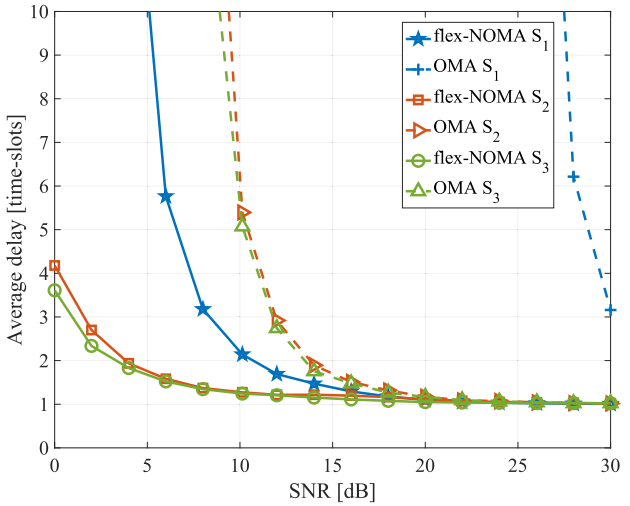

FIGURE 7. Average delay for $K = 4$ and $L_{S_1} = L_{S_2} = L_{S_3} = 10$.

maintain low average end-to-end packet delays. Nonetheless, flex − NOMA enables low-delay transmissions as at each time-slot, scheduling of more than one sources is possible. On the contrary, OMA allows the data of each source to be transmitted at either the $\{S \rightarrow R\}$ or $\{R \rightarrow D\}$ link once every three time-slots and thus, packets remain in the buffer for extended time periods, especially when outages occur for low and medium SNR.

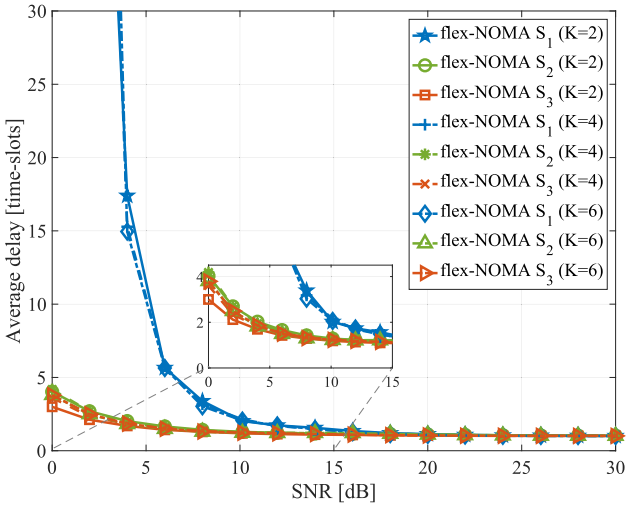

FIGURE 8. Average delay for varying K and $L_{S_1} = L_{S_2} = L_{S_3} = 10$.

The final comparison in Fig. 8 depicts the impact of varying K on the average delay of flex − NOMA. As the cellular user experiences increased outages for low SNR, its packets remain at the relays' buffer for more time-slots compared to the IoT devices. More specifically, SIC at the relays is capable of decoding the low rate signals and since at each time-slot flexible scheduling is allowed, the average delay is low independently of the SNR and converges to one time-slot

after 10 dB. Similarly, as SNR increases, the performance of S_1 improves and its average delay matches the performance of the IoT devices. Finally, it should be noted that although similar delay performance is observed independently of K, as the number of available relays increases, more packets are transmitted in the network, as it has been seen in the outage and sum-rate performance evaluation.

VI. CONCLUSION AND FUTURE DIRECTIONS

A. CONCLUSION

The design of uplink NOMA significantly differs to that of downlink NOMA, especially when multiple relays are available for establishing end-to-end connectivity. In such networks, devices with varying capabilities must be supported, demanding low-complexity network coordination. Moreover, NOMA in the power domain introduces high complexity when performing power allocation, as the existence of multiple relays offers independent wireless channels and thus, achieving asymmetry for interference cancellation might not be feasible at every relay.

In this context, a buffer-aided relay selection policy was proposed, namely flex − NOMA, relying on dynamic decoding ordering at the relays and avoiding the need for channel state information at the transmitter. Furthermore, due to the existence of multiple relays, power allocation at the sources for achieving channel asymmetry, was not employed. On the contrary, all the sources transmitted with equal power and the diversity of the network was exploited, increasing the chances for successful successive interference cancellation. Also, the low-latency demands of 5G services were not neglected, since flex − NOMA aimed at reducing packet delays, based on the use of buffer state information and $\{R \rightarrow D\}$ prioritization.

In addition, a theoretical analysis of flex − NOMA was provided, deriving the link and network outage probabilities. Furthermore, performance evaluation and comparisons with other relevant schemes, in terms of outage probability, average sum-rate and average delay were given. It was shown that flex − NOMA has the potential to accommodate the needs of coexisting users and IoT devices, providing increased robustness and reduced delay, without incurring high implementation complexity.

B. FUTURE DIRECTIONS

While flex − NOMA has promising performance, there are further challenges that need to be investigated. The case of traffic with different priorities should be integrated [34], extending flex − NOMA towards more advanced BSI exploitation.

Moreover, devising pairing algorithms in networks with massive numbers of coexisting users and devices is necessary to improve the performance of NOMA communication [35], since channel asymmetries can be exploited to achieve full SIC after user ordering.

Another important research direction is related to networks where sources are not saturated and random access is facilitated through NOMA [36]. Such scenarios are of increased significance in the context of IoT applications with massive numbers of devices that are sporadically trying to access the channel.

Furthermore, flex − NOMA can be adopted in networks with unmanned aerial vehicles that can be dynamically repositioned [37], [38]. In this way, SIC operation can be improved, compared to fixed topologies.

Finally, another interesting future direction could be related to the enhancements of flex − NOMA to address security issues of 5G-enabled services, including the protection of smart grids.

REFERENCES

[1] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014.

[2] L. Dai, B. Wang, Y. Yuan, S. Han, C.-L. I, and Z. Wang, "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.

[3] S. M. R. Islam, N. Avazov, O. A. Dobre, and K.-S. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 721–742, 2nd Quart., 2017.

[4] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, C.-L. I, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

[5] 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.

[6] M. Shirvanimoghaddam, M. Dohler, and S. J. Johnson, "Massive non-orthogonal multiple access for cellular IoT: Potentials and limitations," *IEEE Commun. Mag.*, vol. 55, no. 9, pp. 55–61, Sep. 2017.

[7] Y. Liu, Z. Qin, M. Elkashlan, Z. Ding, A. Nallanathan, and L. Hanzo, "Nonorthogonal multiple access for 5G and beyond," *Proc. IEEE*, vol. 105, no. 12, pp. 2347–2381, Dec. 2017.

[8] Z. Zhang, H. Sun, and R. Q. Hu, "Downlink and uplink non-orthogonal multiple access in a dense wireless network," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2771–2784, Dec. 2017.

[9] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G non-orthogonal multiple access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6010–6023, Aug. 2016.

[10] S. M. R. Islam, M. Zeng, O. A. Dobre, and K.-S. Kwak, "Resource allocation for downlink NOMA systems: Key techniques and open issues," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 40–47, Apr. 2018.

[11] Y. Gao, B. Xia, K. Xiao, Z. Chen, X. Li, and S. Zhang, "Theoretical analysis of the dynamic decode ordering sic receiver for uplink NOMA systems," *IEEE Commun. Lett.*, vol. 21, no. 10, pp. 2246–2249, Oct. 2017.

[12] J. Wang, B. Xia, K. Xiao, Y. Gao, and S. Ma, "Outage performance analysis for wireless non-orthogonal multiple access systems," *IEEE Access*, vol. 6, pp. 3611–3618, 2018.

[13] 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.

[14] M. Shirvanimoghaddam, M. Condoluci, M. Dohler, and S. J. Johnson, "On the fundamental limits of random non-orthogonal multiple access in cellular massive IoT," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2238–2252, Oct. 2017.

[15] W. Shin, H. Yang, M. Vaezi, J. Lee, and H. V. Poor, "Relay-aided NOMA in uplink cellular networks," *IEEE Signal Process. Lett.*, vol. 24, no. 12, pp. 1842–1846, Dec. 2017.

[16] A. Bletsas, A. Khisti, D. P. Reed, and A. Lippman, "A simple cooperative diversity method based on network path selection," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 3, pp. 659–672, Mar. 2006.
[17] D. Michalopoulos and G. Karagiannidis, "Performance analysis of single relay selection in Rayleigh fading," *IEEE Trans. Wireless Commun.*, vol. 7, no. 10, pp. 3718–3724, Oct. 2008.
[18] Z. Ding, H. Dai, and H. V. Poor, "Relay selection for cooperative NOMA," *IEEE Wireless Commun. Lett.*, vol. 5, no. 4, pp. 416–419, Aug. 2016.
[19] S. Lee, D. B. da Costa, Q.-T. Vien, T. Q. Duong, and R. T. de Sousa, "Non-orthogonal multiple access schemes with partial relay selection," *IET Commun.*, vol. 11, no. 6, pp. 846–854, 2017.
[20] J. Zhao, Z. Ding, P. Fan, Z. Yang, and G. K. Karagiannidis, "Dual relay selection for cooperative NOMA with distributed space time coding," *IEEE Access*, vol. 6, pp. 20440–20450, 2018.
[21] X. Yue, Y. Liu, S. Kang, A. Nallanathan, and Z. Ding, "Spatially random relay selection for full/half-duplex cooperative NOMA networks," *IEEE Trans. Commun.*, vol. 66, no. 8, pp. 3294–3308, Aug. 2018.
[22] A. Ikhlef, D. S. Michalopoulos, and R. Schober, "Max-max relay selection for relays with buffers," *IEEE Trans. Wireless Commun.*, vol. 11, no. 3, pp. 1124–1135, Mar. 2012.
[23] I. Krikidis, T. Charalambous, and J. S. Thompson, "Buffer-aided relay selection for cooperative diversity systems without delay constraints," *IEEE Trans. Wireless Commun.*, vol. 11, no. 5, pp. 1957–1967, May 2012.
[24] Q. Zhang, Z. Liang, Q. Li, and J. Qin, "Buffer-aided non-orthogonal multiple access relaying systems in Rayleigh fading channels," *IEEE Trans. Commun.*, vol. 65, no. 1, pp. 95–106, Jan. 2017.
[25] S. Luo and K. C. Teh, "Adaptive transmission for cooperative NOMA system with buffer-aided relaying," *IEEE Commun. Lett.*, vol. 21, no. 4, pp. 937–940, Apr. 2017.
[26] H. Cao, J. Cai, S. Huang, and Y. Lu, "Online adaptive transmission strategy for buffer-aided cooperative NOMA systems," *IEEE Trans. Mobile Comput.*, vol. 18, no. 5, pp. 1133–1144, May 2019.
[27] N. Nomikos, T. Charalambous, D. Vouyioukas, G. K. Karagiannidis, and R. Wichman, "Relay selection for buffer-aided non-orthogonal multiple access networks," in *Proc. IEEE GLOBECOM Workshops*, Dec. 2017, pp. 1–6.
[28] N. Nomikos, T. Charalambous, D. Vouyioukas, G. K. Karagiannidis, and R. Wichman, "Hybrid NOMA/OMA with buffer-aided relay selection in cooperative networks," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 3, pp. 524–537, Jun. 2019.
[29] M. Oiwa and S. Sugiura, "Reduced-packet-delay generalized buffer-aided relaying protocol: Simultaneous activation of multiple source-to-relay links," *IEEE Access*, vol. 4, pp. 3632–3646, Jun. 2016.
[30] N. Nomikos, T. Charalambous, D. Vouyioukas, and G. K. Karagiannidis, "Low-complexity buffer-aided link selection with outdated CSI and feedback errors," *IEEE Trans. Commun.*, vol. 66, no. 8, pp. 3694–3706, Aug. 2018.
[31] M. Alkhawatrah, Y. Gong, G. Chen, S. Lambotharan, and J. A. Chambers, "Buffer-aided relay selection for cooperative NOMA in the Internet of Things," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 5722–5731, Jun. 2019.
[32] Z. Tian, Y. Gong, G. Chen, and J. A. Chambers, "Buffer-aided relay selection with reduced packet delay in cooperative networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 3, pp. 2567–2575, Mar. 2017.
[33] N. Nomikos, D. Poulimeneas, T. Charalambous, I. Krikidis, D. Vouyioukas, and M. Johansson, "Delay- and diversity-aware buffer-aided relay selection policies in cooperative networks," *IEEE Access*, vol. 6, pp. 73531–73547, Nov. 2018.
[34] M. Vilgelm, H. M. Gürsu, W. Kellerer, and M. Reisslein, "LATMAPA: Load-adaptive throughput- maximizing preamble allocation for prioritization in 5G random access," *IEEE Access*, vol. 5, pp. 1103–1116, Jan. 2017.
[35] J. Zhang, X. Tao, H. Wu, and X. Zhang, "Performance analysis of user pairing in cooperative NOMA networks," *IEEE Access*, vol. 6, pp. 74288–74302, Nov. 2018.
[36] Y. Wu, G. Kang, and N. Zhang, "Random access and resource allocation for the coexistence of NOMA-based and OMA-based M2M communications," *China Commun.*, vol. 14, no. 6, pp. 43–53, 2017.
[37] N. Zhao, X. Pang, Z. Li, Y. Chen, F. Li, Z. Ding, and M.-S. Alouini, "Joint trajectory and precoding optimization for UAV-assisted NOMA networks," *IEEE Trans. Commun.*, vol. 67, no. 5, pp. 3723–3735, May 2019.
[38] A. A. Nasir, H. D. Tuan, T. Q. Duong, and H. V. Poor, "UAV-enabled communication using NOMA," *IEEE Trans. Commun.*, to be published.

NIKOLAOS NOMIKOS (S'12–M'14) received the Diploma degree in electrical engineering and computer technology from the University of Patras, in 2009, and the M.Sc. and Ph.D. degrees from the Information and Communication Systems Engineering Department (ICSD), University of the Aegean, in 2011 and 2014, respectively. He is currently a Postdoctoral Researcher with ICSD, University of the Aegean, and also with the General Department, National and Kapodistrian University of Athens. His research interests include 5G opportunistic mobile networks, cooperative communications, buffer-aided networks, and non-orthogonal multiple access. He is a member of the Technical Chamber of Greece and the IEEE Communications Society (ComSoc).

EMMANOUEL T. MICHAILIDIS was born in Athens, Greece, in 1980. He received the Ph.D. degree in broadband aerospace communications from the University of Piraeus, Piraeus, Greece, in 2011.

Since 2018, he has been an Adjunct Lecturer with the Department of Electrical and Electronics Engineering, University of West Attica, Aigaleo, Greece, and the Department of Informatics and Computer Engineering, School of Engineering, University of West Attica. Since 2012, he has been a Postdoctoral Researcher with the Department of Digital Systems, School of Information and Communication Technologies, University of Piraeus. He has authored or coauthored more than 40 publications in international journals, conference proceedings, and book chapters. He received several best paper awards in his areas of research and his current research interest includes channel modeling and performance analysis of 5G wireless and satellite communication systems.

PANAGIOTIS TRAKADAS received the Dipl.Ing. degree in electrical and computer engineering and the Ph.D. degree from the National Technical University of Athens. He has been with the Technological Educational Institute of Athens as an Assistant Professor, and also with the Technological Educational Institute of Chalkida and the National Technical University of Athens as a Senior Researcher. He is currently an Associate Professor with the National and Kapodistrian University of Athens. He has published more than 80 papers in magazines, journals, books and conferences. He has been actively involved in many research projects. His current research interests include 5G technologies, such as NFV and NOMA, wireless sensor networks, and routing protocols security. He is a Reviewer in several journals and TPC in conferences.

DEMOSTHENES VOUYIOUKAS (S'97–M'04–SM'18) received the five-year Diploma degree in electrical and computer engineering from the National Technical University of Athens (NTUA), in 1996, and the joint M.Sc. degree in engineering economics and the Ph.D. degree in electrical and computer engineering from NTUA, in 2003. He is currently an Associate Professor and the Director of the Computer and Communication Systems Laboratory, Department of Information and Communication Systems Engineering, University of the Aegean, Greece. His research interests include mobile and wireless communication systems, channel characterization and propagation models, performance modeling of wireless networks, cooperative wideband systems with relays, localization techniques, next generation mobile and satellite networks, and MIMO and 5G technologies. In this area, he has more than 110 publications in scientific journals, books, book chapters and international conference proceedings. He is a member of the IEEE Communication Society of the Greek Section, a Member of IFIP and ACM, and also a Member of the Technical Chamber of Greece.

THEODORE ZAHARIADIS received the Dipl.Ing. degree in computer engineering from the University of Patras, Greece, and the Ph.D. degree in electrical and computer engineering from the National Technical University of Athens, Greece. He is currently a Professor with the National and Kapodistrian University of Athens. Since 1997, he has also been a Project Manager or a Technical Manager of many EU funded projects. He has published more than 160 papers in magazines, journals and conferences, and has more than 2300 citations (H-Index 26). His research interests include energy efficiency, smart grids, network virtualization, and wireless sensor networks. He has been a member of Technical Board in multiple scientific conferences and workshops and the Leader Guest Editor in various magazines and journals.

IOANNIS KRIKIDIS received the Diploma degree in computer engineering from the Computer Engineering and Informatics Department (CEID), University of Patras, Greece, in 2000, and the M.Sc. and Ph.D. degrees in electrical engineering from the Ecole Nationale Superieure des Telecommunications (ENST), Paris, France, in 2001 and 2005, respectively. From 2006 to 2007, he was a Postdoctoral Researcher with ENST, Paris, France. From 2007 to 2010, he was a Research Fellow with the School of Engineering and Electronics, University of Edinburgh, Edinburgh, U.K. He is currently an Associate Professor with the Department of Electrical and Computer Engineering, University of Cyprus, Nicosia, Cyprus. His current research interests include wireless communications, cooperative networks, 4G/5G communication systems, wireless powered communications, and secrecy communications.