Signature-based Non-orthogonal Multiple Access (S-NOMA) for Massive Machine-Type Communications in 5G

Mostafa Mohammadkarimi, *Member, IEEE,* Muhammad Ahmad Raza, *Student Member, IEEE,*
Octavia A. Dobre, *Senior Member, IEEE*

Abstract

The problem of providing massive connectivity in Internet-of-Things (IoT) with a limited number of available resources motivates the non-orthogonal multiple access (NOMA) solutions. In this article, we provide a comprehensive review of the signature-based NOMA (S-NOMA) schemes as potential candidates for IoT. The signature in S-NOMA represents the way the data stream of an active device is spread over available resources in a non-orthogonal manner. It can be designed based on device-specific codebook structures, delay patterns, spreading sequences, interleaving patterns, and scrambling sequences. Additionally, we present the detection algorithms employed to decode each device's data from non-orthogonally superimposed signals at the receiver. The bit error rate of different S-NOMA schemes is simulated in impulsive noise environments, which can be important in machine-type communications. Simulation results show that the performance of the S-NOMA schemes degrades under such conditions. Finally, research challenges in S-NOMA oriented IoT are presented.

I. Introduction

With the explosive demand for higher data rate, last decades have witnessed the evolution of the mobile wireless communication systems from the first generation to the fourth generation (4G). However, the objective of the future wireless communication systems (5G and beyond) will not only be to provide services with higher data rate. One of the main goals of these systems is to support massive machine-type communications (mMTC) within the paradigm of the Internet-of-Things (IoT). In mMTC a large number of devices are connected to one base station (BS), each transmitting sporadically a small payload information without any human interactions. According to a forecast by Ericson, several billions of IoT devices will be installed by 2020 [1].

To support the ubiquitous connectivity, new advancements of communication technologies and protocols are required. One of the fundamental aspects in this regard is the medium access control (MAC). Since the number of available resources, such as frequency and time, are limited, conventional orthogonal multiple access (OMA) schemes cannot support the massive connectivity in IoT. The large gap between the available resources and required massive connections motivates the research community to investigate non-orthogonal multiple access (NOMA) solutions. In this regard, several schemes have been considered by the 3GPP New Radio standard and the low power wide area (LPWA) technologies [2]–[4].

M. Mohammadkarimi is with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada. (e-mail: mostafa.mohammadkarimi@ualberta.ca). M. A. Raza, and O. A. Dobre are with the Department of Electrical and Computer Engineering, Memorial University, St. John's, NL, Canada (e-mail: odobre@mun.ca and maraza@mun.ca).

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), through its Discovery program.

The signature-based NOMA (S-NOMA) represents a class of schemes in which the data from each active device is spread over available resources in a non-orthogonal manner by using a set of predefined signatures. This provides additional degrees of freedom (DoF) for device separation, which leads to a significantly high overloading factor (OLF).[1] The signatures can be designed based on device-specific codebook structures, delay patterns, spreading sequences, interleaving patterns, and scrambling sequences. At the BS, advanced activity detection and multi-user detection (MUD) algorithms are employed to decode each device's data from the non-orthogonally superimposed signals. The joint design of signatures and MUD receiver is performed to improve the resilience of each device's signal against multiuser interference (MUI) and to obtain a low-complexity MUD receiver [5]. Fig. 1 presents a taxonomy of the S-NOMA schemes. In addition to the classification based on signature, each scheme can be implemented following either a grant-free (GF) or request-grant (RG) approach.

This article provides a comprehensive review of different S-NOMA schemes, which can support massive uplink connectivity in IoT. Both GF and RG S-NOMA scenarios are discussed. Moreover, the bit error rate (BER) of various S-NOMA schemes is investigated through simulation in the presence of impulsive noise. Finally, potential research challenges for S-NOMA in IoT are presented.

II. S-NOMA: Random Access (RA) Procedure

The conventional RG- and GF-based uplink multiple access schemes cannot support massive connectivity in machine-type communications since the maximum number of supported devices is limited by the number and scheduling granularity of orthogonal resources. A promising way to scale up the uplink connectivity in IoT applications is to use S-NOMA in both RG- and GF-based MAC schemes. A brief presentation of the conventional random access (RA) procedure, as well as of the RG and GF S-NOMA, is provided in the following.

A. Collision in the Conventional RA Procedure

Fig. 2a shows the four-step handshake RA procedure in the conventional RG-based multiple access employed in Long-Term Evolution (LTE). In the RA procedure, during the RA channel (RACH) stage, more than one device can select the same orthogonal preambles. Identical preamble selection leads to preamble collision and makes this RG-based multiple access scheme ineffective for massive uplink connectivity. Similarly, the conventional GF-based multiple access schemes employed in wireless networks, such as ALOHA, slotted ALOHA, and their variants suffer from low throughput due to the high probability of collision in massive connectivity. Furthermore, in both RG- and GF-based schemes, successful data transmission is achieved through OMA; however, there is a limited number of orthogonal resources to be allocated to a very large number of devices

B. Request-grant S-NOMA

The main idea behind RG S-NOMA is to reduce the probability of collision and increase the OLF in the RACH process of the RA procedure by employing advanced MUD. Recently, an RG S-NOMA scheme was proposed in [6], where the BS utilizes the difference in the time-of-arrival (ToA) to identify multiple devices with identical preambles, as shown in Fig. 2b. This

[1]The OLF is defined as the ratio of the number of overloaded signals to the number of orthogonal resources.

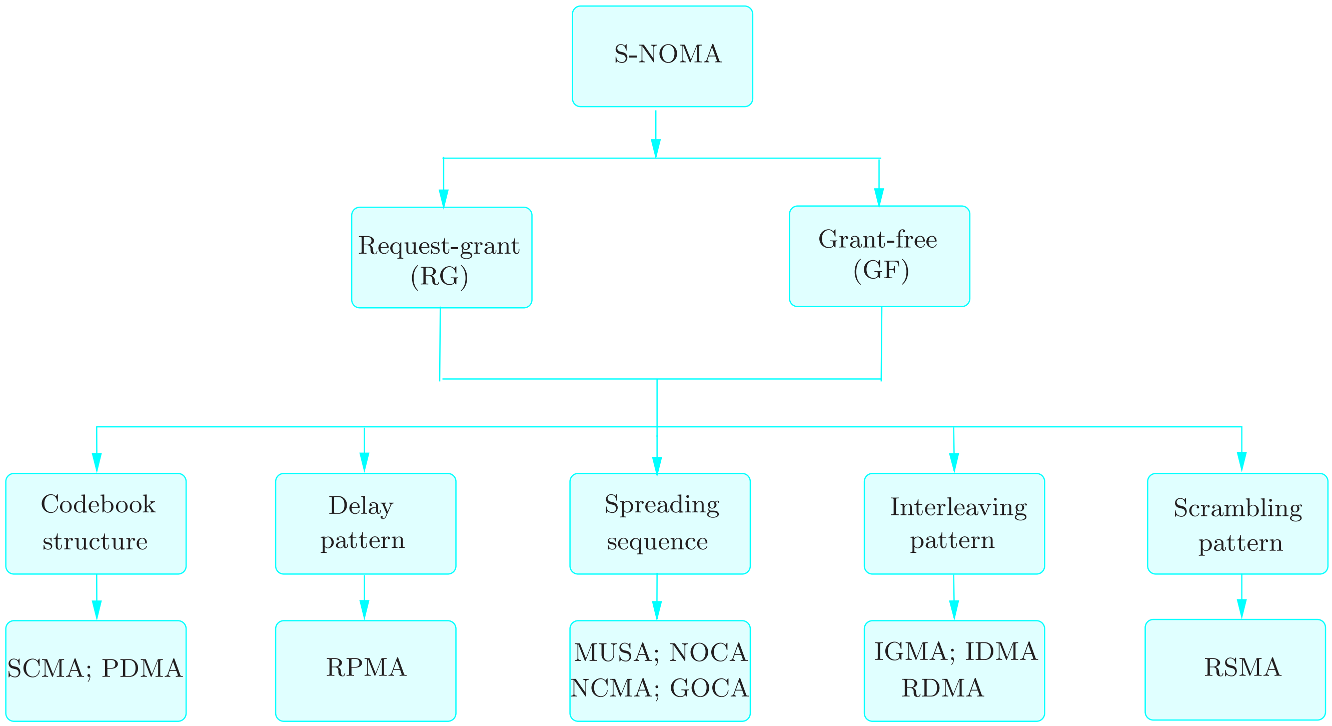
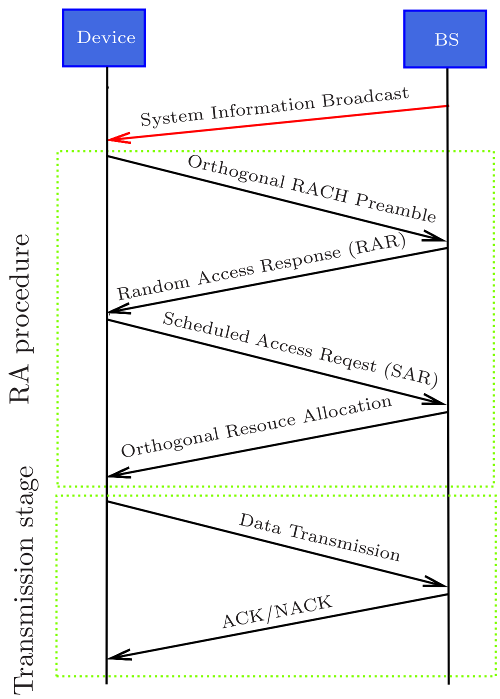


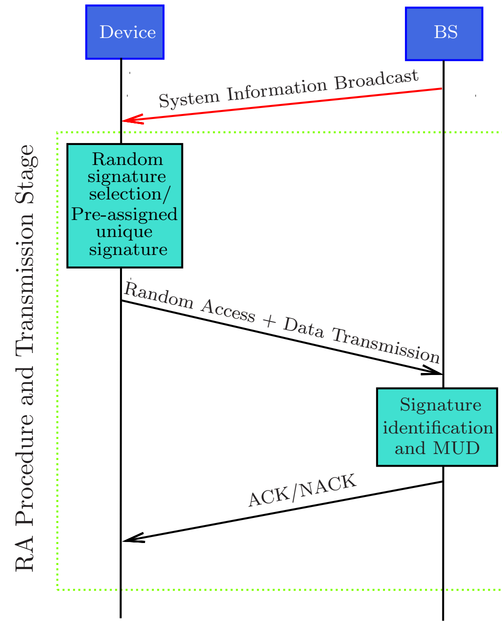
Fig. 1: Taxonomy of S-NOMA schemes. SCMA: sparse code multiple access; PDMA: pattern division multiple access; RPMA: random phase multiple access, MUSA: multi-user shared access, NOCA: non-orthogonal coded access, NCMA: non-orthogonal coded multiple access, GOCA: group orthogonal coded access; IGMA: interleave-grid multiple access, IDMA: interleave-division multiple-access, RDMA: repetition division multiple access; RSMA: resource spread multiple access.

represents a modification of the LTE RA procedure, by including a delay pattern S-NOMA scheme in the RACH process. As seen, after preamble transmission, the BS implements multi-preamble detection to detect the delayed versions of the transmitted preambles. Then, it considers the devices with detected collisions, i.e., devices with identical preamble, as a non-orthogonal random access (NORA) group and responds to the group with a random access response (RAR) message. Based on the channel conditions obtained through preamble detection, the RAR message includes the power back-off values for devices in the NORA group. Each device estimates its distance from the BS based on the received signal strength of cell-specific reference signal to obtain its ToA. Based on the estimated ToA, each device searches the closest ToA value in the RAR message. The power back-off information is utilized for the power-domain multiplexing at the device-side for scheduled access request (SAR), as well as for performing successive interference cancellation (SIC) at the BS. A more general approach to NORA grouping is to directly employ the S-NOMA signatures at the RACH stage to further reduce the probability of collision and increase the OLF.

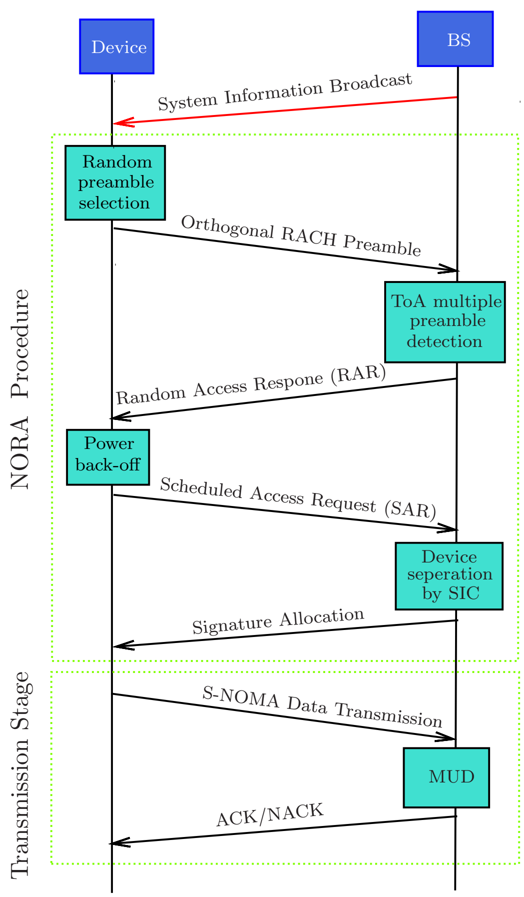
Finally, in the transmission stage, the BS allocates resources to those devices that successfully transmitted SAR. As distinct signatures are allocated to devices in this stage, collisions do not occur at the expense of higher control signaling and higher complexity MUD in the RA procedure when compared with the GF S-NOMA scheme; this will be described in the next section.



(**a**) Conventional RG OMA in the LTE [6].



(**c**) GF S-NOMA [7].



(**b**) RG S-NOMA (A modification of the LTE channel access) [6].

Fig. 2: RA procedure and transmission stage in a) conventional RG OMA [6] b) RG S-NOMA [6] and c) GF S-NOMA [7]. ACK: acknowledgment; NACK: negative acknowledgment.

C. Grant-free S-NOMA

In grant-free S-NOMA, each active device directly transmits its packets to the base station (BS) without waiting for any permission. Fig. 2c shows the joint RA procedure and transmission stage in GF S-NOMA. Both random signature selection and pre-assigned signature allocation schemes have been investigated for GF S-NOMA [7]. In the former scheme, each device randomly selects a signature from a large pool for link establishment. On the other hand, under the latter scheme, each device is pre-assigned a unique signature used for all transmission slots. This signature thus serves as the device identification, as well. These schemes increase the spectral efficiency and number of served devices through the simultaneous non-orthogonal RACH and data transmission at devices, as well as joint signature identification, and detection of device activity and data at the BS. GF S-NOMA allows asynchronous, non-orthogonal contention-based access that is well suited for sporadic uplink transmissions of small data bursts common in IoT applications. In order to support a large number of devices, the network can be partitioned into groups, with a signature pool allocated to each group.

III. Signature Design in S-NOMA

The S-NOMA OLF and complexity of the MUD receiver at the BS depend on the design of signatures. To design efficient signatures, different operations, such as linear spreading, multi-dimensional modulation, interleaving, and scrambling can be employed [5]. In this section, we overview how these operations are employed at bit-level and/or symbol-level to provide efficient signatures.

A. Codebook Structure S-NOMA

The main idea behind the codebook structure S-NOMA is the direct mapping of each device's data stream onto a codeword in a structured codebook. The sparsity pattern used in codebooks is different in order to facilitate device separation and reduce MUI, while it is kept the same in codewords belonging to the same codebook. The sparsity pattern of codebooks defines the signatures, and the number of non-zero elements in a codeword represents the number of orthogonal resources used for transmission; this equals the diversity order. The spreading factor (SF) is equivalent to the length of a codeword, and the OLF is determined by the ratio of the number of multiplexed devices in a group to the spreading factor. Two potential schemes in this category are sparse code multiple access (SCMA) and pattern division multiple access (PDMA). Fig. 3 (dashed box) illustrates the codebook structure S-NOMA in which every two bits are directly mapped into a codeword with length four.

1) SCMA provides the same diversity order for different devices. Its implementation is similar to LTE, with a key difference of the joint design of modulation and low-density spreading. This design is implemented through a multi-dimensional modulation with lower number of projection points. A multi-dimensional modulation is simply the mapping of the input coded bits onto the points in the multiple complex dimensions (containing both I and Q). By employing multi-dimensional modulation, the SCMA codewords provide both diversity and shaping gain [8].

2) PDMA sparsely maps each devices's data stream onto a group of resources according to a PDMA codebook to realize disparate transmission diversity order. A resource group can consist of time, frequency, and spatial resources, or any combination of them. Data of multiple devices can be multiplexed onto the same resource group with a different PDMA pattern. In general,

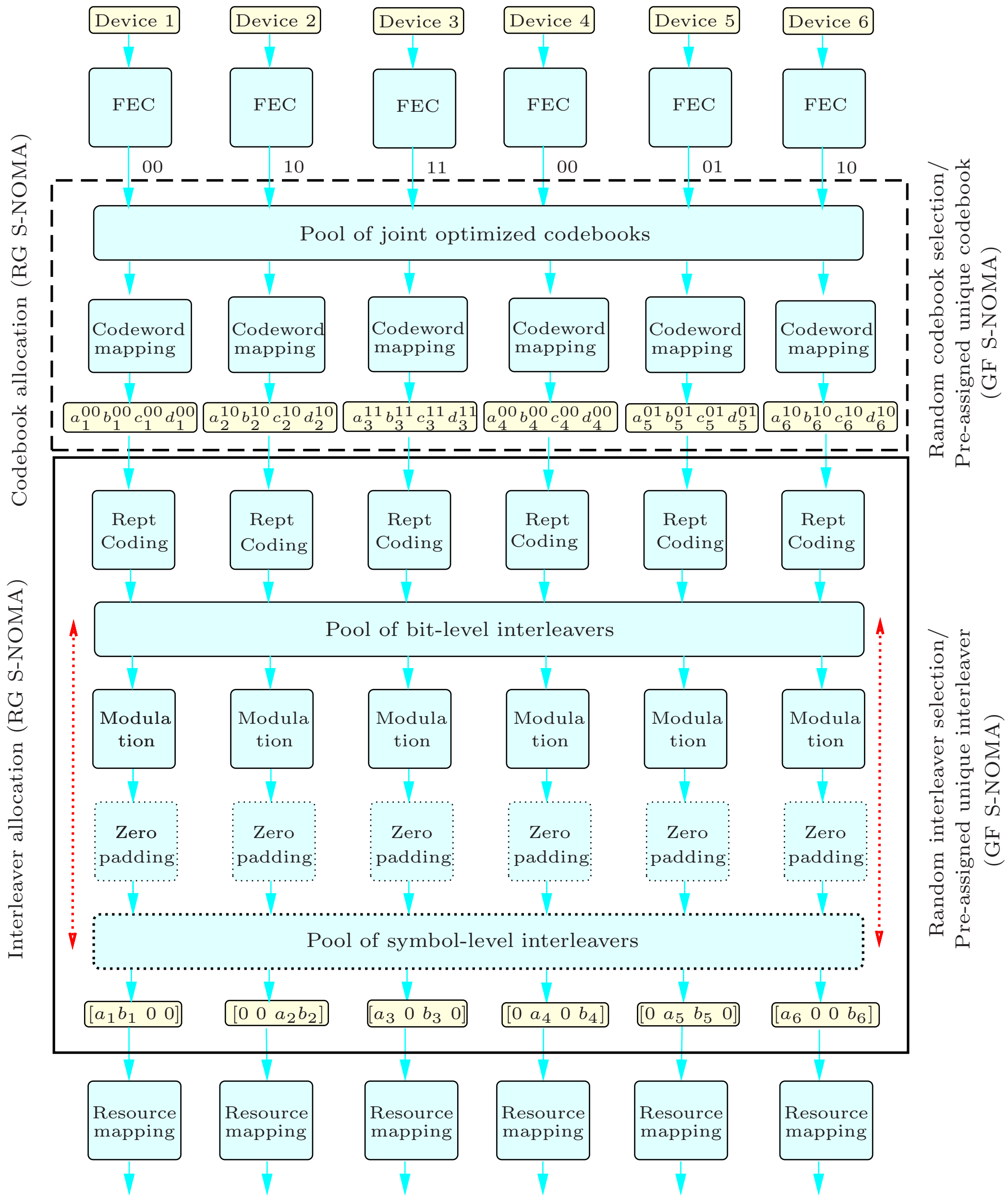


Fig. 3: Codebook structure S-NOMA with SF 4 and OLF 150% (Dashed box). Every two bits are directly mapped into a codeword with length 4 in a codebook. Interleaving pattern S-NOMA with SF 4 and OLF 150% (Solid box). The codewords consist of two zero entries through zero-padding and a_i and b_i, which are the modulated symbols of the ith device. Rept stands for repetition.

if N is the size of the orthogonal resource group, there are $2^N - 1$ possible binary vectors for the transmission pattern. Assuming K is the number of devices, we can thus choose K transmission pattern vectors out of the $2^N - 1$ candidates to construct codebooks. The transmission patterns are chosen to maximize the diversity order while minimizing the overlaps among devices. The same transmission diversity order can be considered in PDMA while minimizing the maximum inner product between pattern vectors. After transmission pattern selection, all codewords in a codebook are designed following the same transmission pattern vector [9].

B. Interleaving Pattern S-NOMA

The key characteristic of interleaving S-NOMA is that different interleavers along with repetition and/or low-rate forward error correction (FEC) coding are employed for device separation. The length of the repetition code and/or the inverse of code rate determines the SF, while the interleaver patterns define the signatures. Superposition transmission with different interleaving patterns results in interference averaging, and thus, lower MUI. Three schemes in this category are: interleave-grid multiple access (IGMA), interleave-division multiple access (IDMA), and repetition division multiple access (RDMA). Fig. 3 (solid box) illustrates the interleaving pattern S-NOMA concept.

1) IGMA provides high DoF for device separation by employing different bit-level interleavers, different symbol-level grid mapping patterns, and a combination of bit-level interleavers and grid mapping patterns. In IGMA, low code-rate FEC coding and repetition coding can be considered as a spreading technique. After channel coding and repetition, the bit-level interleaver is employed to make the transmission bits randomly distributed. Then, in the grid mapping process, sparse mapping based on zero padding and symbol-level interleaving are introduced to provide another dimension for device multiplexing. Sparse mapping leads to further MUI reduction, and the symbol-level interleaving randomizes the symbol sequence order, which may further bring benefits in terms of combating frequency-selective fading and inter-channel interference. IGMA can achieve coding and diversity gains [2].

2) IDMA is a special case of IGMA, where different devices are distinguished by bit-level interleavers. The bit-level interleaver, along with repetition and/or very low-rate FEC coding, enables dispersion of data across a long bit stream, which can provide time-frequency diversity. The main advantages of IDMA consist of high OLF and low BER when it uses very low-rate FEC. Also, the interleaved low-rate codes with a simple chip-by-chip iterative decoding strategy could achieve the capacity of a Gaussian multiple access channel [10].

3) RDMA employs different simple cyclic-shift repetition patterns in the symbol-level to design device-specific signatures. The cyclic-shift repetition behaves as a randomizer (interleaver), and provides both time and frequency diversity. Moreover, the repetition patterns lead to randomized MUI in both time and frequency domain for each repeated modulated symbol. On the contrary to IDMA, random interleaver is not used in RDMA [11].

C. Spreading Sequence S-NOMA

The idea behind spreading sequence S-NOMA is that low cross-correlation real/complex-valued spreading sequences are employed as device-specific signatures in order to enable non-orthogonal transmission[2]. Since the spreading sequences are not necessarily orthogonal, this leads to high DoF for device separation [5]. After linear spreading, symbols are directly mapped to orthogonal resources. While both low-density spreading (LDS)- and non-LDS sequences can be considered, the former is preferable since it can efficiently mitigate MUI. The OLF is defined the ratio of the number of devices to the length of the spreading sequence. Existing S-NOMA schemes in this category are: multi-user shared access (MUSA), non-orthogonal coded access (NOCA), non-orthogonal coded multiple access (NCMA), and group orthogonal coded access (GOCA). The main

[2]Complex sequences provide additional DoF for device separation.

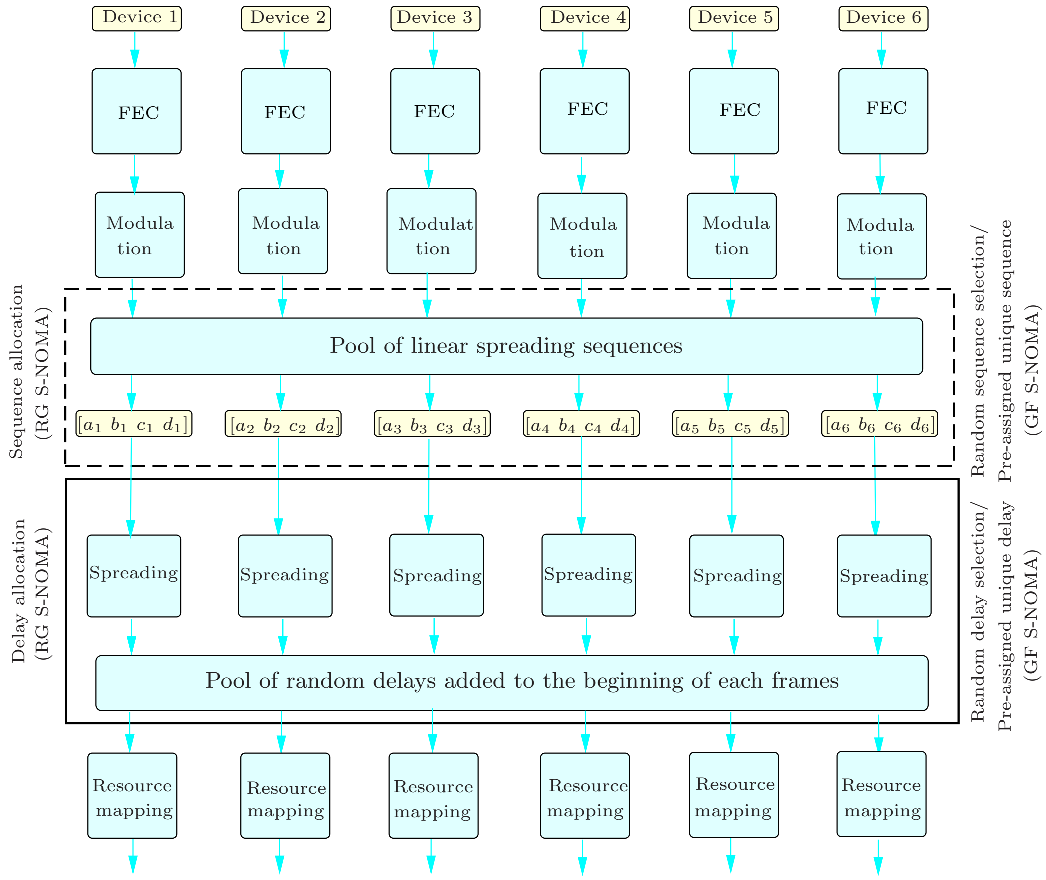


Fig. 4: Spreading sequence S-NOMA with SF 4 and OLF 150% (Dashed box). The entries of the ith codeword, i.e., $[a_i\ b_i\ c_i\ d_i]$, represent the product of a modulated symbol and the spreading sequence of the ith device. Delay pattern S-NOMA (Solid box); devices employ identical spreading sequences, such as the Gold code.

difference between these schemes is related to the spreading sequences used for device separation. In Fig. 4 (dashed box), the concept of the spreading sequence S-NOMA is illustrated.

1) MUSA: Each device's modulated data symbols are spread by a special family of low cross-correlation complex-valued spreading sequences, which can facilitate MUD based on SIC implementation. The real and imaginary parts of the employed non-binary spreading sequence are drawn from an M-level value set ($M \geqslant 2$) with uniform distribution. As an example, $\{-1, 0, 1\}$ is a popular set for MUSA, leading to 9^L sequences, with L as the sequence length. The main advantage of MUSA consists of its high OLF; this can be as high as 700% in multipath fading channel. Moreover, the blind detection based on the SIC receiver makes MUSA an efficient GF S-NOMA scheme [2].

2) NOCA: employs the non-LDS LTE-defined low cross-correlation sequences for symbol-level spreading. Frequency-domain and/or time-domain spreading can be achieved through resource mapping. The former leads to multicarrier (MC)-code division multiple access (CDMA), and the latter results in MC-direct sequence-CDMA. NOCA can be developed with small changes in the LTE standard. Also, it is effective to establish a quasi-synchronous channel such that the BER is almost the same as for the synchronous case, when the uplink synchronization is kept within one cyclic prefix [2].

3) NCMA: Grassmannian sequences are employed as user specific sequences in NCMA, and aim at maximizing the minimum chordal distance between codeword pairs. Grassmannian sequences represents one kind of Welch-bound equality sequences;

hence, they are the optimal sequences maximizing the ergodic sum capacity for equal received power. However, the elements of the Grassmannian sequences are irregular complex values, which may increase the complexity for hardware implementation [2].

4) GOCA is considered as a long sequence-based S-NOMA scheme, which exploits a dual spreading sequence. The spreading sequences in GOCA have a two-stage structure, which enables group separation through a set of non-orthogonal sequences and device separation within a group through a set of orthogonal sequences. GOCA is an enhanced version of RDMA. The main difference between GOCA and RDMA is that the former employs orthogonal sequences to spread the modulation symbols into shared time and frequency resources after repetition coding in each group. By employing the SIC receiver, GOCA can support a high OLF due to the large DoF of the non-orthogonal sequences and a significant MUI reduction due to the orthogonal sequences in the same group [11].

D. Delay Pattern S-NOMA

The main idea behind the delay pattern S-NOMA is to separate devices in the delay domain. In this scheme, devices with identical linear spreading sequence are distinguished from their delay at the BS.[3] It should be mentioned that this scheme is usually employed in GF S-NOMA and in the NORA procedure of RG S-NOMA. Fig. 4 (solid box) illustrates the concept of delay pattern S-NOMA.

The random phase multiple access (RPMA) is one of the schemes in this category. It differs from the CDMA in two ways: firstly, all devices in a group employ the same Gold code to spread the bits before transmission; secondly, the transmissions are not synchronized and start with a random delay. RPMA supports dynamic SF selection without the intervene of the BS. A dynamic SF is important in order to minimize the power consumption and enable adaptive transmission. For example, when the channel condition is unfavourable, a higher SF can be used to reduce the power consumption. On the other hand, when the channel condition permits, a lower SF can be used to transmit at higher rates [3].

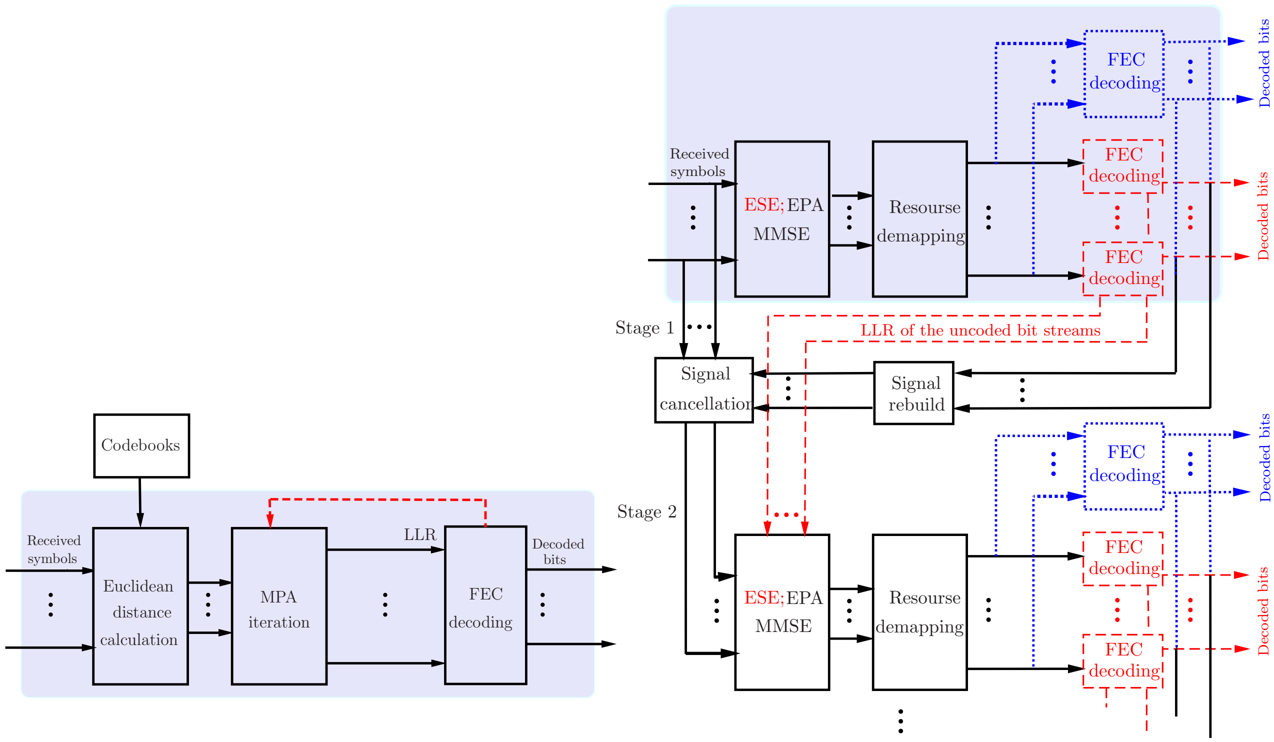
E. Scrambling Pattern S-NOMA

The main characteristic of the scrambling pattern S-NOMA is that large DoF for device separation can be obtained through the combination of very low-rate FEC and device-specific scrambling sequences. The low-rate FEC codes result in high coding gain, and the long low cross-correlation scrambling sequences reduce the MUI. Similar to interleaving pattern S-NOMA, the inverse of the code rate represents the SF. A scheme in this category is resource spread multiple access (RSMA).

RSMA employs the combination of very low-rate FEC and long scrambling sequences designed for uplink wideband CDMA (and optionally different interleavers) to spread a group of different devices' signals over the entire orthogonal resources assigned to the group. Spreading of bits to the entire resources results in high diversity order, which enables decoding at negative signal-to-interference-plus-noise ratio (SINR) [2].

By changing the resource mapping configuration, single-carrier (SC)-RSMA and MC-RSMA can be achieved. The former is optimized for low-power consumption and coverage extension for short packet transmission. Additionally, it employs low

[3]Correct decoding is possible as long as two transmission frames do not arrive at the exact same moment.



(a) MPA receiver.

(b) SIC (dot blue lines) and PIC (red dash lines) receivers.

Fig. 5: Block diagrams of MUD receivers for S-NOMA.

peak-to-average power ratio (PAPR) modulations and enables asynchronous access. On the other hand, the latter is optimized for low-latency access [2].

IV. MUD in S-NOMA

The employed receivers in the uplink S-NOMA are categorized as: message passing algorithm (MPA), expectation propagation algorithm (EPA) SIC, MPA-SIC, parallel interference cancellation (PIC), and pseudo-noise (PN) array despreader [3], [12].

A. MPA Receiver

The optimum receiver for S-NOMA is the joint maximum a posteriori probability (MAP) receiver, which gives zero probability of error in the absence of noise and does not suffer from the near-far problem. However, it exhibits high computational complexity. In order to achieve a performance close to that of the MAP receiver with significantly lower computational complexity, the iterative MPA receiver can be employed (Fig. 5a) for the S-NOMA schemes with LDS signatures, such as SCMA and PDMA. This MUD receiver jointly decodes all devices by employing graphical models, which capture the kinds of independence and factorization structure due to the sparse structure of the signatures to reduce the complexity of the joint MAP receiver.

The decoding in the MPA receiver is performed on the Tanner graph constructed by the codebook, where a function node (FN) represents an orthogonal resource and a variable node (VN) represents the symbol of a device. The inputs to the MAP receivers are: the received signal on each orthogonal resource, channel estimate on each resource from each device, and noise power estimate on each resource. The decoding procedure starts with the initial conditional probability calculation at each FN. Then, it enters message passing iterations between FNs and VNs along the edges. There are two steps for each iteration, referred to as FN update and VN update, which are done independently for each FN-VN pair. After enough iterations, the probability guess of each codeword is obtained at a VN and then is employed for log-likelihood-ratio (LLR) calculation. Finally, the achieved LLRs are directly input to the turbo decoder. The MPA receiver requires no constraint on the SINR difference among devices at the receiver, and is robust to the channel correlation between devices. However, its complexity is higher than that of SIC and PIC receivers [12].

B. EPA Receiver

This receiver is a graphical-based MUD algorithm. Compared to the MAP receiver, the EPA receiver approximates the sophisticated posterior distributions in the message update steps with simple distributions through distribution projection by using the Kullback-Leibler divergence. The EPA receiver is used in S-NOMA schemes with sparse signature, such as SCMA, to archive near MAP performance with lower computational complexity. By employing the EPA receiver, the complexity order of SCMA decoding is reduced to linear complexity, i.e., it only scales linearly with the codebook size and the average degree of the FNs in the factor graph representation [13].

C. SIC Receiver

This is the simplest, lowest complexity receiver for S-NOMA, and is particularly appropriate for schemes based on spreading sequence and scrambling pattern, such as MUSA, NOCA, NCMA, GOCA, and RSMA. The SIC receiver decodes devices one after another by treating all the other undecoded devices as interference when decoding one target device. At each stage, one or multiple devices with highest SINR are decoded. Then, the decoded signals are reconstructed and cancelled out in the next stage (dot blue lines in Fig. (5b)). The SIC receiver exhibits low BER when the SINR levels of devices are significantly different. However, its performance degrades when this SINR difference is trivial, because of the error propagation. Hence, power back-off mechanisms are required to provide different received powers at the BS. This is a major challenge in GF S-NOMA schemes. In order to improve the performance of the SIC receiver, linear receivers, such as minimum mean square error (MMSE) are employed during each stage of the SIC receiver [12]. Recently, EPA-SIC has also been developed.

D. MPA-SIC Receiver

The combination of the MPA and SIC receivers benefits of their advantages in terms of near-optimal performance and low-complexity, respectively, and is employed in S-NOMA with LDS signatures, such as SCMA and PDMA.

In such a receiver, the MPA is first applied to a limited number of devices in order of decreasing power, so that the number of colliding devices over each orthogonal resources does not exceed a given threshold value, K_f. Then, the successfully decoded devices by the MPA are removed by SIC and the procedure continues until all devices are successfully decoded or no device gets successfully decoded by MPA. Since MPA is used for a limited number of devices instead of all devices in each stage,

Table I: Features of S-NOMA schemes.

	Attributes	SCMA	MUSA	IDMA	PDMA	IGMA	RPMA	NOCA	RSMA
1	Proposed by	Huawei	ZTE	Nokia	CATT	Samsung	Ingenu	Nokia	Qualcomm
2	Signature	Codebooks	Spreading sequence	Bit-level interleaver	Codebooks	Bit- and symbol-level interleavers	Delay	Spreading sequence	Symbol-level scrambler
3	Sparse signature	Yes	No	No	Yes	Yes	No	No	No
4	MUD receiver	MPA(-SIC) EPA-(SIC)	MMSE-SIC	ESE-PIC	MPA MPA-SIC	MPA ESE-PIC	PN array despreader	MMSE-SIC	MMSE-SIC MF-SIC
5	Receiver complexity	High Moderate	Low	Moderate	High Moderate	High Moderate	High	Low Moderate	Low
6	Device overload	Moderate	High	High	Moderate	High	High	Moderate	High

the decoding complexity is significantly reduced. It is worth noting that the SIC-MPA receiver becomes MPA when K_f is equal to the number of active devices, and it reduces to the SIC receiver when $K_f = 1$ [12].

E. PIC Receiver

Opposite to the SIC receiver, the PIC receiver decodes devices in parallel. At each stage, all devices are decoded and the results are fed back as prior information to the next stage (red dash lines in Fig. (5b)). In general, PIC exhibits less delay when compared with SIC, while its complexity is higher. Linear MUD receivers, such as MMSE and elementary signal estimator (ESE) can be employed at each stage of the PIC receiver to improve the performance [12].

ESE-PIC is a soft-cancellation-based receiver, mostly used by S-NOMA schemes with bit-level operations like IDMA and IGMA, to provide robust performance for high OLF. It first detects transmitted symbols via ESE detection; then, the detected symbols are deinterleaved and decoded in parallel to acquire coding gain. The output information from the decoder is finally sent back to the ESE to aid symbol detection [10].

F. PN Array Despreader

This receiver is employed for S-NOMA schemes based on delay patterns, such as RPMA. Since the BS has no knowledge of the intentional delays that the devices add to the beginning of the frame in GF S-NOMA, a PN array despreader is employed to despread all the received waveforms through brute-force operation for all possible chip arrival times and SFs. Each chip hypothesis is despread, deinterleaved, decoded, and then checked via a cyclic redundancy check (CRC) to determine the valid frames [3].

Summary: Table I summarizes the features of the discussed S-NOMA schemes. In particular, the spreading signature, MUD receiver, complexity of receiver, and OLF are provided.

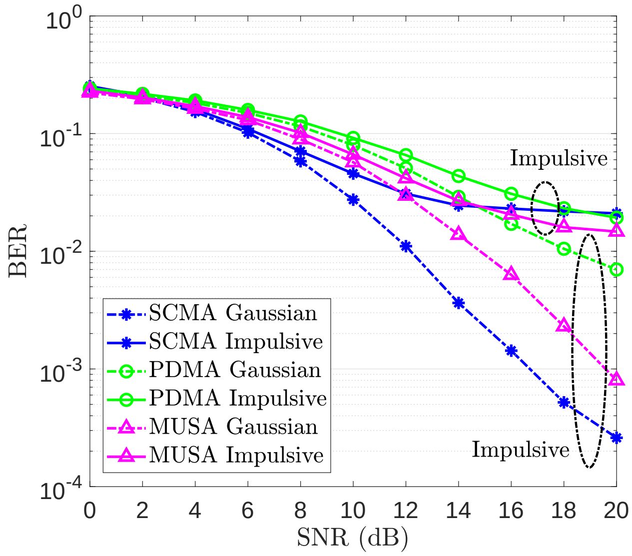


Fig. 6: Performance degradation of the S-NOMA schemes in the presence of impulsive noise.

V. Performance of the S-NOMA Schemes in Impulsive Noise

In S-NOMA, when the received superimposed signal is impaired by impulsive noise in time domain or by a narrow-band interferer/tone-jamming in frequency domain, some of the time domain chips or some of the subcarriers in the frequency domain may be completely corrupted, which results in performance degradation.

In this section, we evaluate the performance of the S-NOMA schemes in terms of BER. We characterize the impulsive noise by employing the Bernoulli-Gaussian model [14], with the probability density function $f_{\mathsf{w}}(x) = (1-\epsilon) f_{\mathrm{g}_1}(x) + \epsilon f_{\mathrm{g}_2}(x)$, where $0 < \epsilon < 1$, and $f_{\mathrm{g}_1}(x)$ and $f_{\mathrm{g}_2}(x)$ are zero-mean complex Gaussian distributions with variances $\sigma^2_{\mathrm{g}_1}$ and $\sigma^2_{\mathrm{g}_2}$. We set $\epsilon = 0.01$ and $\sigma^2_{\mathrm{g}_2}/\sigma^2_{\mathrm{g}_1} = 100$.

Fig. 6 illustrates the BER of the SCMA, PDMA, and MUSA, employing quadrature phase-shift-keying modulation in flat Rayleigh fading, while being contaminated by Gaussian and impulsive noise at the receiver. The OLF is considered 150%, while the number of devices and orthogonal resources are 6 and 4, respectively. The SCMA codebooks and PDMA patterns in [15] are employed. The real and imaginary parts of the complex spreading sequences in MUSA are uniformly drawn from $\{-1, 0, 1\}$. As seen, the BER of the S-NOMA schemes degrade in the presence of impulsive noise when compared with Gaussian noise. Moreover, the performance degradation in SCMA and PDMA is higher than in MUSA since the MPA receiver is more sensitive to non-Gaussianity than the MMSE-SIC receiver.

VI. Challenges and Future Trends

There are several issues that need to be addressed in uplink S-NOMA. Some of these challenges are as follows:

Joint signature and MUD receiver design in non-Gaussian noise: The performance of the S-NOMA schemes in their current forms degrades in the presence of non-Gaussian/impulsive noise. Hence, design of new signatures and MUD receiver are required to address this problem.

Activity detection error: The BER performance of the MUD receivers degrades in the presence of the signature identification error in GF S-NOMA. It is important to do a quantitative study of this degradation, e.g., due to the false alarm rate. To the best of the authors' knowledge, such analysis is not available in the literature.

OLF in NORA: The performance of the RG S-NOMA scheme in [6] may be improved by employing RPMA in the RA process. Opposite to RPMA, the transmission delay in [6] is not intentionally added by devices; rather, it is due to the propagation delay. Hence, the DoF for device separation are restricted by the positions of the devices in the IoT network. Using RPMA can provide additional DoF.

Channel estimation: The performance of the MUD algorithms in the S-NOMA schemes heavily relies on the efficiency of channel estimation. New channel estimation methods for S-NOMA are required to be developed.

MUD receiver complexity: For the MPA receiver, the complexity may still be high for massive connectivity. Therefore, new MUD receivers should be designed to reduce complexity. One approach is using the Gaussian approximation for interference-plus-noise; this approximation tends to be more accurate as the connectivity becomes larger in IoT.

VII. CONCLUSION

Uplink S-NOMA schemes for massive IoT were reviewed in this article. The modifications of GF and RG schemes were introduced for S-NOMA. For different S-NOMA schemes, the design of signatures was presented, along with the transmitter and receiver structures, and receiver complexity and performance. Moreover, the performance of the SCMA, PDMA, and MUSA schemes in the presence of impulsive noise was studied through simulation, and it was shown that they are generally sensitive to impulsive noise. Finally, various directions for future investigation were provided.

REFERENCES

[1] 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.

[2] S. Yang, P. Chen, L. Liang, J. Zhu, and X. She, "Uplink multiple access schemes for 5G: A survey," *ZTE Communications*, vol. 15, no. S1, pp. 31–40, Jun. 2017.

[3] T. J. Myers, "Random phase multiple access communication interface system and method," Aug. 24 2010, US Patent 7,782,926.

[4] K. Chandra, A. S. Marcano, S. Mumtaz, R. V. Prasad, and H. L. Christiansen, "Unveiling capacity gains in ultradense networks: Using mm-Wave NOMA," *IEEE Veh. Technol. Mag.*, vol. 13, no. 2, pp. 75–83, Apr. 2018.

[5] F.-L. Luo and C. Zhang, *Signal Processing for 5G Algorithms and Implementation*. John Wiley, 2016.

[6] Y. Liang, X. Li, J. Zhang, and Z. Ding, "Non-orthogonal random access (NORA) for 5G networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4817–4831, Jul. 2017.

[7] R1-164269, "Contention-based non-orthogonal multiple access for UL mMTC," *3GPP*, May 2016.

[8] F. Wei and W. Chen, "Low complexity iterative receiver design for sparse code multiple access," *IEEE Trans. Commun.*, vol. 65, no. 2, p. 621–634, Feb. 2017.

[9] S. Chen, B. Ren, Q. Gao, S. Kang, S. Sun, and K. Niu, "Pattern division multiple access—a novel nonorthogonal multiple access for fifth-generation radio networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 4, pp. 3185–3196, May 2017.

[10] L. Ping, L. Liu, K. Wu, and W. K. Leung, "Interleave division multiple-access," *IEEE Trans. Wireless Commun.*, vol. 5, no. 4, pp. 938–947, Apr. 2006.

[11] R1-167535, "New uplink non-orthogonal multiple access schemes for NR," *3GPP*, May 2016.

[12] R1-166098, "Discussion on the feasibility of advanced MU-detector," *3GPP*, Aug. 2016.

[13] X. Meng, Y. Wu, Y. Chen, and M. Cheng, “Low complexity receiver for uplink SCMA system via expectation propagation,” in *Proc. IEEE WCNC*, Mar. 2017, pp. 1–5.

[14] R. Pighi, M. Franceschini, G. Ferrari, and R. Raheli, “Fundamental performance limits of communications systems impaired by impulse noise,” *IEEE Trans. Commun.*, vol. 57, no. 1, pp. 171–182, Jan. 2009.

[15] B. Wang, K. Wang, Z. Lu, T. Xie, and J. Quan, “Comparison study of non-orthogonal multiple access schemes for 5G,” in *Proc. IEEE BMSB*, Jun. 2015, pp. 1–5.