IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 7 November 2022, accepted 3 January 2023, date of publication 10 January 2023, date of current version 23 January 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3235825

RESEARCH ARTICLE

Design of Incidence Matrices With Limited Constellation Expansion in Massive Connectivity NOMA Systems

ELI HWANG, XING HAO, (Member, IEEE), AND GUILLERMO E. ATKIN, (Life Senior Member, IEEE)
Department of Electrical and Computer Engineering, Illinois Institute of Technology, Chicago, IL 60616, USA
Corresponding author: Eli Hwang (ehwang2@hawk.iit.edu)

ABSTRACT Non-orthogonal multiple-access (NOMA) is designed to transmit massive amounts of user communications. The incidence matrix manages the relationship between users and resources. This study focused on increasing user supportability and complexity reduction using larger incidence matrices. Our approach to optimizing the incidence matrix to improve system capacity and reduce complexity is based on two critical mathematical concepts: Parallel classes in hypergraph theory and combinatorial designs allow us to explore and extend incidence matrices. Frame theory is used to estimate matrix structures. Then, we investigate applications utilizing our incidence matrix designs–Simple Orthogonal Multi-Arrays (SOMA). The characteristics of SOMA reflect the unique Latin Square pattern, allowing us to produce a highly flexible and fair resource allocation matrix. SOMA designs let us support overload factors over 500%. The theoretical performance analysis equations of our NOMA system were established to support dynamic adaptability and optimization. We implemented and evaluated security methods for eavesdroppers. The prototype of the user hierarchy allows a higher-priority group to have a lower error rate without significantly affecting the system's performance. Finally, the Monte Carlo simulation indicated that our NOMA systems allow higher degrees of freedom and lower complexity than other NOMA schemes while maintaining graceful error rates with a maximum 33% improvement.

INDEX TERMS Incidence matrix, NOMA, combinatorial design, frame theory, SOMA.

I. INTRODUCTION

This study explores the design of large incidence matrices used in non-orthogonal multiple-access (NOMA) by combinatorial design techniques for a given number of users b or resources v to achieve better bit error rates (BER), throughput, and overload capability (OC).

There are three main reasons why it is challenging to generate and apply large incidence matrices to NOMA systems.

- The linear combination of modulated symbols
- The mapping of constellations with uncertainty
- The excessive complexity of decoding

Therefore, the complexity of constellation expansion limits the size of the incidence matrices used by many sparse NOMA schemes. Consequently, we were motivated to fill a void in generating and analyzing larger incidence matrices for massive user support.

Through a combinatorial design known as Simple Orthogonal Multi-Arrays (SOMA), our method is capable of producing larger dimension incidence matrices that are confirmed to be near equiangular tight frames (ETFs) by frame theory. ETFs have significant applications in signal processing and coding theory because of their robustness to noise and transmission losses [1]. ETFs are characterized by the fact that the coherence between any two distinct row or column vectors is equal to the Welch bound [2]. This guarantees that the maximum coherence between pairs of vectors is minimized. Therefore, our incidence matrix designs can be used comprehensively in other NOMA designs with suboptimized user resource allocation.

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

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

We chose this optimization because the lower the coherence, the less the multiuser signal overlaps [3], and hence, the less complex the constellation expands. Therefore, we can replace the complex computation for resource and user allocation of each specific NOMA scheme by providing an existing structure to obtain a suboptimal result.

NOMA is envisioned to be a promising multiple-access technique for future wireless communication systems. Its purpose is to meet user experience needs more efficiently and provide a theoretical basis for the later deployment of cells with smaller coverage [4], [5]. NOMA schemes use nonorthogonality and overloading to achieve their purpose by serving multiple users in the same orthogonal resource element (RE). While these concepts are not new in 3G or 4G communications, such as CDMA [6], NOMA has a non-orthogonal approach at the core of its design structure, aiming for a larger OC. Therefore, one of the most critical aspects of NOMA is the handling of the overloaded architecture, that is, the incidence matrix.

Many NOMA schemes have been widely investigated in recent years. Newer and more complex systems beyond the initial power and code domain types, such as hybrids [7] and waveforms [8], are being discussed. The more complex and extensive the system, the more it needs to be organized under a particular mathematical guideline to maintain its efficiency. Therefore, optimizing incidence matrices, particularly higher-dimensional matrices, has become increasingly important because an optimized incidence matrix can reduce the complexity of the receiver and maximize performance [9].

The effect of optimizing the incidence matrix can be observed in pattern-division multiple access (PDMA) [10]. By optimizing the power scaling and phase-shifting factors in its pattern matrix, PDMA systems have significantly improved the spectral efficiency over OMA systems. Moreover, the author in [9] demonstrated that the regularity of regular sparse NOMA contributes to optimizing the spectral efficiency by employing user and resource mapping over the NOMA with an irregular matrix. In addition, the incidence matrix following the ETF structures allocates the resource block selection for each user in a manner that reduces the number of maximum likelihood calculations in the optimal receiver for each user.

Additionally, the incidence matrices are where the NOMA system manages the user and resource distribution and regulates the system size and overload factor. This fact motivated us to study high-dimensional incidence matrices. On the other hand, it is well known that ongoing research is trying to find an optimized scheduler in the power domain NOMA, such as [11] and [12]. However, the incidence matrix, which is the counterpart of the scheduler in code domain NOMA, has received little attention. One of the reasons is that, although the incidence matrix performs similar tasks to the scheduler in terms of user grouping and resource distribution, it also has additional responsibilities, including taking part in the transmission, encoding, and decoding of multiuser data, which makes it more complex. As a result, most studies used only a low-dimension 4×6 matrix as an example because of its complexity. Our design method can reduce the computational complexity of generating the incidence matrix and decoding at the receiver, which is discussed in Section II-E.

In this study, we propose an incidence matrix structure that utilizes a Resolvable Balanced incomplete block design (BIBD) in combinatorial design. Resolvability is an essential property in hypergraph theory that divides the incidence matrix into parallel classes to deal with constellation expansion problems [13]. Frame theory was used to support and examine the matrix structures. We also explored and expanded the boundaries of NOMA applications based on the characteristics of our incidence matrix design.

A. RELATED WORKS

As stated in the previous introduction, there has been little attention in the literature on creating larger incidence matrices for NOMA schemes. However, it is well known that a matrix with a longer girth performs better in the message passing algorithm (MPA), one of the major multiuser detections (MUD) methods. Therefore, we can still find related work that attempts to optimize the incidence matrix.

In [14], incidence matrices based on multistage maximum distance separable (MDS) codes were generated. The author traded performance with complexity, enabling the system to be scalable. The largest matrix in [14] was 56×70.

In [15], the author exploited an algebraic scheme to improve the low-density spreading (LDS) sequence design based on projective geometry. Owing to the expansion of the matrix size, the performance was better than that of other LDS designs. The largest matrix in [15] was 13×15.

Finally, the author in [16] utilized the Latin rectangle to improve the power efficiency of a high-dimensional codebook of sparse code multiple access (SCMA). The largest matrix in [16] was 8×12.

Comparatively, our design is more adaptable than the above methods because it is based on parallel classes with resolvable characteristics, allowing methods to expand or puncture the matrix. Frame theory also proves that the structure yields a type of optimal packing of lines in Euclidean space. Hence, our matrix design was not limited to a specific NOMA scheme. The largest matrix we provide in this study is 20×100, as shown in Figure 6.

B. CONTRIBUTIONS

In this study, we combined a code domain NOMA (CD-NOMA) system established from previous studies [17], [18] with SOMA to take advantage of its highly flexible structure. In our system, different SOMAs can provide NOMAs with incidence matrices of varying sizes. The freedom of this design structure provides an advantage in achieving optimum trade-offs between performance, OC, and security by changing the incidence matrix. Thus, we achieved more extensive connectivity and higher throughput than SCMA [19], PDMA, Multi-User Shared Access (MUSA) [20], and

our prior work. This study makes the following main contributions:

- We begin by utilizing SOMA designs to form an incidence matrix generating algorithm. Examples of incidence matrices with OCs of 166% and 500% are provided.
- We verified the efficiency of the generated incidence matrix using frame theory. We can facilitate the design of large incidence matrices, particularly for massive communications and low-capacity channels.
- We provide an objective function to optimize the incidence matrix based on performance parameters.
- Computational complexity is calculated to show that our design has low complexity.
- Hereafter, we show how to deal with the growing problem of mapping between patterns and constellations using an expurgation. Expansion methods for the incidence matrix are also discussed.
- We provide applications that utilize the characteristics of SOMA, like resolvable and BIBD.
- We demonstrate that, by using our MUD, we can use well-known modulation performance analysis equations to predict our system. Owing to the adaptability of our system, we provide a fundamental step towards resilience via resource allocation.
- We performed simulations for both additive white Gaussian noise (AWGN) and Rayleigh fading channels. Hereafter, we evaluate the numerical models to compare their system performance with those of other NOMA schemes. We found that our system performed better than well-defined NOMA schemes.

The remaining sections of this paper are organized as follows. In Section II, we discuss the incidence matrix design in our current and previous work. Frame theory, which can describe the characteristics of our design, was also introduced. The application of the characteristics of the incidence matrix is explained in Section III. The simulation results and discussion are presented in Section IV. Finally, conclusions are presented in Section V.

II. INCIDENCE MATRIX DESIGN

Combinatorial design theory is a mathematical tool that challenges questions regarding whether it is possible to arrange elements of a finite set into patterned incidences, such as sets, words, and arrays, so that a particular balance rule can be satisfied [21].

We exploited the most researched combinatorial design, balanced incomplete block design (BIBD), to generate the incidence matrix for NOMA because the sparse pattern of BIBD can provide an inherently sparse and evenly distributed mapping to its resources and users in a non-orthogonal manner [22].

We introduce our previous works that utilized a specific type of BIBD and explain why we upgraded our work. A new structure was used to enhance security and cover flaws in the old design.

Hereafter, we introduce another mathematical tool called Frame Theory to show that our design structure is optimized for OC, user fairness, and regularity for a set of parameters, such as the number of resources and users [23]. In addition, both our previous and new designs can be shown to be near ETFs. An ETF is the closest equivalent of orthonormal bases in incoherence [24], meaning that the incidence matrix has the maximal spatial angular displacement for every column vector pair, making the structure optimal for detecting multiuser signals.

Finally, we discuss methods to adjust the matrix parameters and list the objective functions to optimize the OC of the incidence matrix under the constraints of power, resources, and BER.

A. PREVIOUS WORKS

In our previous studies [17] and [18], we studied a Low-Density Code design to build a CD-NOMA system by taking advantage of highly structured BIBDs. We further propose a method to achieve a larger OC for transmitting more users with limited resources by expurgating the number of users per resource to obtain a structure with larger sparsity.

Many communication systems use BIBDs, including sensor networks, modulation schemes, and Low-Density Parity Check codes (LDPCs) [25]. BIBD is a design that selects all sets of the same size. Additionally, every treatment occurred equally and the design was balanced pairwise. This balance property enables us to translate BIBDs into regular incidence matrices [26] by treating each block as a column (user) and each element as a row (resource). Resources can be viewed as many things, such as power, sparse codes, and spaces.

Definition 1 (BIBD [22]): A BIBD is a pair (X, A) defined by D(v, b, r, k), where X is a set of v elements called points, and A is a collection of subsets of X called blocks, such that

- $|X| = v$
- Every block contains exactly k points
- Each element of X appears exactly in r points out of b blocks
- Every pair of distinct points is contained in exactly b blocks.

Hence, by treating each block as a user and each element as a resource, we can create an arbitrary incidence matrix from the BIBD array D(v, b, r, k), as shown in Figure 1.

RE1	1	1	…	…	r	0
RE2	1	0	…	…	…	0
…	k	…	…	…	…	…
REv	0	0	…	…	…	1
	UE1	UE2	…	…	…	UEb

FIGURE 1. An example of arbitrary incidence matrix D(v, b, r, k).

- **total b users (columns) and total v frequencies (rows)**
- **r users per resource and k resources per user**
- **Their relation is given by $k \cdot b = r \cdot v$**

In our design, various BIBDs can provide NOMA with various coding matrices. However, we focused on a specific BIBD called the Steiner Triple System (STS). One of the main reasons why we chose STS is that almost all STS are resolvable [27]. A design is resolvable if one can partition its blocks into parallel classes, where a set of blocks contains every element exactly once. Utilizing the resolvable property, we can compress the incidence matrix in a balanced manner, which helps us group and detect multiuser information. Another reason is that STS structures can be processed as Steiner ETFs [28], which is the natural choice when one tries to combine the advantages of orthonormal bases with the concept of redundancy provided by frames [29].

Definition 2 (STS) [22]: Given three integers t, k, v such that $2 < t < k < v$, a Steiner system S(t, k, v) is a v-set X together with a family A of k-subsets of X (blocks), with the property that every t-subset of X is contained in exactly one block.

An example of the STS structure in NOMA utilizing the resolvable property is shown below. The 4 color codes represent the 4 parallel classes based on the resolvable property.

$$\mathrm{STS}(9) = \begin{Bmatrix} (1,2,3),(4,5,6),(7,8,9), \\ (1,4,7),(2,5,8),(3,6,9), \\ (1,5,9),(2,6,7),(3,4,8), \\ (1,6,8),(2,4,9),(3,5,7), \end{Bmatrix}$$

We can translate each pair like (1,2,3) into a user with resources allocated to 1, 2, and 3 to generate a 9×12 incidence matrix, where a column is a user and a row is a resource.

$$\mathrm{STS}(9) = \mathrm{D}(\mathrm{v}=9, \mathrm{b}=12, \mathrm{r}=4, \mathrm{k}=3) =$$

$$\begin{bmatrix} 1&0&0&1&0&0&1&0&0&1&0&0 \\ 1&0&0&0&1&0&0&1&0&0&1&0 \\ 1&0&0&0&0&1&0&0&1&0&0&1 \\ 0&1&0&1&0&0&0&0&1&0&1&0 \\ 0&1&0&0&1&0&1&0&0&0&0&1 \\ 0&1&0&0&0&1&0&1&0&1&0&0 \\ 0&0&1&1&0&0&0&1&0&0&0&1 \\ 0&0&1&0&1&0&0&0&1&1&0&0 \\ 0&0&1&0&0&1&1&0&0&0&1&0 \end{bmatrix}$$

$$= \begin{bmatrix} u_1&u_4&u_7&u_{10} \\ u_1&u_5&u_8&u_{11} \\ u_1&u_6&u_9&u_{12} \\ u_2&u_4&u_9&u_{11} \\ u_2&u_5&u_7&u_{12} \\ u_2&u_6&u_8&u_{10} \\ u_3&u_4&u_8&u_{12} \\ u_3&u_5&u_9&u_{10} \\ u_3&u_6&u_7&u_{11} \end{bmatrix}$$

As we can see, the incidence matrix can be compressed into dimension $v \times p$, where p is the number of parallel classes, which helps us generate new complex symbols by combining multiuser data.

However, while this structure provides much larger user support, its k points are always 3, and the STS pattern is limited. As mentioned in [17] and [18], the receiver's performance can be improved by using low-density matrices with a more extended k, which is not possible by STS. In addition, a limited pattern needs to be secured. After introducing the SOMA structure, we proved a size comparison in the following subsection. In addition, our SOMA designs have smaller constellation sizes per RE than the STS structures. Therefore, we can achieve better restrictions on constellation expansion by utilizing the updated structure, which enhances the motivation to update our structure reference.

B. SOMA DESIGN

We exploited the SOMA structure [30], [31] to provide a general framework for producing larger incidence matrices. Through frame theory, it is known that SOMA structures generally distribute larger, if not the maximum, spatial angular displacements in their column vectors.

The characteristics of SOMA ensure an inherent, sparse, and consistent weighted array, which is a crucial property of the incidence matrix. These features make SOMA a perfect tool for designing an incidence matrix.

Therefore, the aims of our SOMA design are as follows:

- Increase efficiency and performance compared to other systems.
- A layer of security is provided using the natural structure of the Latin Squares (LS) during the design phase [32].

Compared with the STS structure, SOMA has more parallel classes with more resolution options. This increased degree of freedom will grant us higher security than our previous design owing to the wider variety. For example, SOMA(2,4) shown below can be grouped into parallel classes, coded in colors, by selecting either rows, columns, or transversals [33].

In rows:

1,5	2,6	3,7	4,8
2,7	1,8	4,5	3,6
3,8	4,7	1,6	2,5
4,6	3,5	2,8	1,7

In columns:

1,5	2,6	3,7	4,8
2,7	1,8	4,5	3,6
3,8	4,7	1,6	2,5
4,6	3,5	2,8	1,7

In transversals:

1,5	2,6	3,7	4,8
2,7	1,8	4,5	3,6
3,8	4,7	1,6	2,5
4,6	3,5	2,8	1,7

SOMA also has a flexible k value (resource per user) other than a fixed value of 3 by STS. The SOMA structure is constructed from Latin Squares, which have applications in

cryptography because of their variability [32], [34]. Hence, the SOMA patterns inherit diversity to prevent the adversary from quickly guessing the incidence matrix.

Consequently, we propose a method for building an incidence matrix based on SOMA that is composed of Mutually Orthogonal Latin Squares (MOLS). Although detailed definitions of SOMA and MOLS are found in [31] and [35], we provide a simplified description of orthogonal LS, MOLS, and SOMA below.

Definition 3: (Orthogonal Latin Square): A Latin Square consists of n rows and n columns with n symbols with values from 1 to n. Each row and column will contain symbols from 1 to n exactly once. We call this $n \times n$ array an order n LS, or LS(n). Suppose there is a pair of LS, A, and B, of the same order n such that the superposition of their elements $(a_{ij}, b_{ij}) \neq (a_{kl}, b_{kl})$ if $(i, j) \neq (k, l)$, where $i, j, k, l \in \{1 \sim n\}$, then we call the pair of LS A and B orthogonal LS. The matrix obtained by the superposition of A on B is called a Greco–Latin or Euler square.

Definition 4: (Mutually Orthogonal Latin Square:) A set of MOLS is a set of two or more LS of the same order, all of which are orthogonal to one another.

For example, 3 different MOLS(6) are shown in Figure 2a.

MOLS1

2	3	4	5	1
3	4	5	1	2
4	5	1	2	3
5	1	2	3	4
1	2	3	4	5

MOLS2

3	5	2	4	1
4	1	3	5	2
5	2	4	1	3
1	3	5	2	4
2	4	1	3	5

MOLS3

4	2	5	3	1
5	3	1	4	2
1	4	2	5	3
2	5	3	1	4
3	1	4	2	5

(a)

2,3,4	3,5,2	4,2,5	5,4,3	1,1,1
3,4,5	4,1,3	5,3,1	1,5,4	2,2,2
4,5,1	5,2,4	1,4,2	2,1,5	3,3,3
5,1,2	1,3,5	2,5,3	3,2,1	4,4,4
1,2,3	2,4,1	3,1,4	4,3,2	5,5,5

(b)

FIGURE 2. (a) Example of 3 different MOLS(6) (b) Superimposition of 3 MOLS(6).

An example of a MOLS is shown in Figure 2a. and b., we confirmed that none of the cells contained a repeated set of elements for any two or more colors combinations [26].

Another property of MOLS is that the number of MOLSs of an order is bounded by $t = n - 1$. Thus, we call a set of t-MOLS(n) a complete set, if $t = n - 1$. If we let Num(n) denote the size of the largest collection of MOLS(n), we obtain the following characteristics [35]:

1) $Num(n) \leq n - 1 \text{foranyn} \geq 2$
2) If q is a prime power, then $Num(q) = q - 1$
3) $Num(n) \geq 2$ for all n except 2 and 6

Definition 5 (Simple Orthogonal Multi-Array): Let $k > 0$ and $n > 1$. SOMA(k, n) is an $n \times n$ array whose cells each contain exactly k elements chosen from a given set of $k \cdot n$ symbols, such that each of the symbols occurs exactly once in each row and once in each column, together in at most one cell. Note that a SOMA(1,n) is the same as a Latin Square of order n (LS(n)) and that a set of k superimposed Mutually Orthogonal Latin Squares (MOLS) of order n gives a SOMA(k, n) [31].

We can see from the definitions above that the challenge of our design lies in finding the optimized order of the MOLS to balance efficiency, performance, and system complexity, which we discuss in a later subsection. However, the generation of a single SOMA was simple. While researchers already find many MOLS, column and row switching can be applied to create different SOMA to avoid exposing the incidence matrix to the unauthorized user. Hence, a generalized algorithm that generates an incidence matrix utilizing a known MOLS database was proposed.

Algorithm 1 SOMA(k, n) Design

Inputs: Number of resources v, users b, and resources per user k

Output: Incidence matrix SOMA(k, n)

1. Find the k - MOLS(n) by b where $b = n^2$ from a pre-generated combinatorial database.
2. Ordered the sequence of the chosen MOLS.
3. Extend the element values by the sequence order.
4. Perform random column and row switching on all MOLS for security.
5. Superimpose the combination of MOLS to generate SOMA(k, n).
6. Translate the SOMA(k, n) to incidence matrix D(v, b, k, r) by treating each cell as a user and each element as a resource.

While Latin Square is not a BIBD structure, it can be exploited to generate one structure. Using similar techniques, we can translate the SOMA design into an incidence matrix following the format of BIBD, D(v, b, r, k), by arranging each cell of SOMA as a user and the number of elements as the REs of the user.

Therefore, an arbitrary SOMA(k, n) array can represent a system that supports total n^2 users (columns) and $k \cdot n$ REs (rows). Hence, the relationship between arbitrary SOMA(k, n) and D(v, b, r, k) is $v = k \cdot n$; $b = n^2$; $r = n$; and $k = k$;

To give a SOMA structure example, we show SOMA (3, 5) in Figure 3, which is composed of superimposed MOLS1, MOLS2, and MOLS3 with their elements rearranged according to SOMA's definition. The difference between Figure 2b. superimposition of the 3 MOLS(6), and

2,8,14	3,10,12	4,7,15	5,9,13	1,6,11
3,9,15	4,6,13	5,8,11	1,10,14	2,7,12
4,10,11	5,7,14	1,9,12	2,6,15	3,8,13
5,6,12	1,8,15	2,10,13	3,7,11	4,9,14
1,7,13	2,9,11	3,6,14	4,8,12	5,10,15

FIGURE 3. An example SOMA(3, 5).

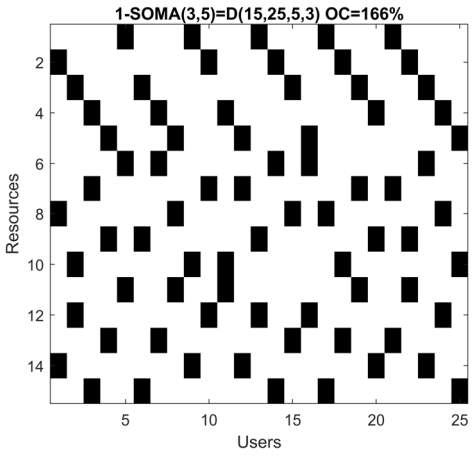


FIGURE 4. An intensity map of SOMA(3,5).

Figure 3. SOMA(3, 5), are the elements of SOMA assigned in sequence.

This SOMA(k, n), where $k = 3$ and $n = 5$, represents a system with $n^2 = 25$ users (column) and $k \cdot n = 15$ frequencies (row). By arranging each cell of the SOMA as a user and the number of elements as frequencies, we can translate this array into an incidence matrix of format D($v = 15$, $b = 25$, $r = 5$, $k = 3$). The corresponding incidence matrix of SOMA(3, 5) is shown in Figure 16, as shown at the bottom of the page, and its sparsity is shown in Figure 4. From Figure 4, we can visually observe that the ratio of the sparsity of SOMA(3,5) is $\frac{1}{5}$, which originates from the order $n = 5$ of this SOMA structure. Therefore, the higher the order of SOMA, the sparser the incidence matrix becomes. The growth in sparsity can help us to aid sparse types of NOMA schemes, as in [36].

Finally, we would like to prove that SOMA designs have smaller constellation sizes per RE than STS designs when supporting the same number of users. We know that the constellation size is controlled by parameter r in BIBD(v, b, r, k), where r represents the number of users per RE. STS(v) has a relationship of $v \cdot r = b \cdot 3$ and $r = \frac{3\cdot(v-1)}{6}$, while SOMA($k$, n) has its $r = n$ and $b = n^2$. Therefore, when the number of supported users is the same in both systems, r in SOMA(k, n) becomes $r_{SOMA} = \sqrt{\frac{v(v-1)}{6}}$. Comparing r_{SOMA} and r_{STS}, we find that if $v \to 4$, $r_{SOMA} = \sqrt{2} < r_{STS} = \frac{3}{2}$, and if $v \to 5$, $r_{SOMA} = \sqrt{\frac{5(4)}{6}} < r_{STS} = 2$. Assume $v \to N$, $r_{SOMA} < r_{STS}$, then $v \to N+1$, $r_{SOMA} = \sqrt{\frac{(n+1)n}{6}} < r_{STS} = \frac{n}{2}$. By Mathematical induction, r in SOMA(k, n) will always be smaller than r in STS when both designs support the same number of users. Thus, the expansion of the constellation generated by the SOMA design will be slower than that of STS, which can help us to provide higher user support.

C. FRAME THEORY

We used Frame Theory to show the characteristics of our design. Frame theory focuses on studying overcomplete representations of orthonormal bases, which fits the concept of nonorthogonality in NOMA schemes [37]. The theorems and conclusions were employed to improve our understanding of the mapping problem [38]. Consequently, we leverage frame-theoretic constructs to analyze the limits of the generic CD-NOMA model.

The STS and SOMA designs can be expanded as EFTs, meaning that the matrix is the closest to uniformly orthogonal between each column vector, so the calculation complexity for multiuser detection can be reduced.

If we treat the arbitrary incidence matrix D(v, b, r, k) defined in Figure 1. as a frame F that belongs to space $C^{v\times b}$ in frame theory, and assign f_i as a column vector of F. Then, frame F can be seen as a synthesis matrix that compresses higher-dimensional information into a lower dimension.

$$\text{SOMA}\,(3,5) = \text{D}\,(15,25,5,3) = \begin{vmatrix} 0&0&0&0&1&0&0&0&1&0&0&0&1&0&0&0&1&0&0&0&1&0&0&0&0\\ 1&0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&1&0&0&0&1&0&0&0\\ 0&1&0&0&0&1&0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&1&0&0\\ 0&0&1&0&0&0&1&0&0&0&1&0&0&0&0&0&0&0&0&1&0&0&0&1&0\\ 0&0&0&1&0&0&0&1&0&0&0&1&0&0&0&1&0&0&0&0&0&0&0&0&1\\ 0&0&0&0&1&0&1&0&0&0&0&0&1&0&1&0&0&0&0&0&1&0&0&0&0\\ 0&0&1&0&0&0&0&0&1&0&1&0&0&0&0&0&1&0&1&0&0&0&0&1&0\\ 1&0&0&0&0&0&0&1&0&0&0&0&0&1&0&1&0&0&0&0&0&1&0&0&0\\ 0&0&0&1&0&1&0&0&0&0&0&1&0&0&0&0&0&1&0&1&0&0&0&0&1\\ 0&1&0&0&0&0&0&0&1&0&1&0&0&0&0&0&1&0&0&0&0&0&1&0&0\\ 0&0&0&0&1&0&0&1&0&0&1&0&0&0&0&0&0&0&1&0&0&1&0&0&0\\ 0&1&0&0&0&0&0&0&0&1&0&0&1&0&0&1&0&0&0&0&0&0&0&1&0\\ 0&0&0&1&0&0&1&0&0&0&0&0&0&1&0&0&1&0&0&1&0&0&0&0&0\\ 1&0&0&0&0&0&0&1&0&0&1&0&0&0&0&0&0&1&0&0&1&0&0&1&0\\ 0&0&1&0&0&1&0&0&0&0&0&0&1&0&0&1&0&0&0&0&0&0&1&0&1 \end{vmatrix} \tag{1}$$

F^H performs a reverse operation [39].

$$F \triangleq C^b C^v \tag{2}$$

$$F^H \triangleq C^v C^b \tag{3}$$

Moreover, we know that a matrix $F \in C^{v \times b}$ with $v < b$ is a frame over the finite Hilbert space C^v if its linearly dependent column vector $f_i \in C^v$ satisfies the frame bounds equation [23]:

$$\alpha \|z\|_2^2 \leq \sum_{i=1}^{b} |\langle f_i z \rangle|^2 \leq \beta \|z\|_2^2, \forall z \in C^v \tag{4}$$

where z is any column vector that belongs to space C^v, and the finite coefficients $0 < \alpha < \beta$ are referred to as the highest lower and lowest upper frame bounds for α and β, respectively.

With this information, we can measure parameters critical to frame theory, such as frame redundancy, unit-normality, tightness, and incoherence.

Frame redundancy, P(F) (5), measures the expansion in mapping, where OC originates from [23] and [39]. P(F) is key to most frame applications in engineering. For example, redundancy reduces the inevitable quantization error that appears in all applications of series representations [5].

$$P(F) \triangleq \frac{b}{v} \tag{5}$$

Based on the above definition, the frame redundancy measures the over-completeness of frame F. In other words, overloading the original signal space C^v by b frame vectors represents the expected robustness gained under the synthesis matrix operation [40].

Unit-normality, defined as $\|f_j\|_2$, outlines fairness in terms of the energy distribution between column vectors.

Tightness (6) is used to describe the energy profile of F:

$$FP(F) \triangleq \left\| F^H F \right\|_F^2 \geq \frac{b^2}{v} \tag{6}$$

Frame Potential (FP) measures the representative energy dispersed by a normalized frame F for vectors over a unit multidimension.

A frame is said to be tight when both bounds of inequality in (4) are tight.

$$\alpha = \beta = P(F) \triangleq \frac{b}{v} \tag{7}$$

This condition guarantees that the global minimum of (6) can be determined. Therefore, we can obtain the lower Welch bounds (WBs) via (a) the Welch bound is equivalent to the frame potential inequality in frame theory and (b) the frame potential is minimized at tight frames [40], [41]. WBs are important tools for spreading in the vector space. Its mathematical form is shown in (8).

$$\left(\left| \max_{i \neq j} |\langle f_i, f/_j \rangle| \right| \right)^{2t} \geq \frac{1}{b-1} \left(\frac{b}{\binom{v+t-1}{t}} - 1 \right), t \in Z^+ \tag{8}$$

Tight frame conditions can also help us perform a perfect reconstruction from the synthesis domain to the analysis domain via the frame operator S_F [23]:

$$S_F \triangleq FF^H \tag{9}$$

Incoherence (10) measures the regularity of F, which has been considered in other NOMA schemes [42].

$$\mu(F) \triangleq \max_{j \neq i} \frac{\left| (f_j \cdot f_i) \right|}{\|f_j\|_2 \cdot \|f_i\|_2} \tag{10}$$

Through incoherence, we determine the highest off-diagonal entry of the frame Gramian matrix [43].

$$G_F \triangleq F^H F \tag{11}$$

The frame Gramian operator, G_F, satisfies the following properties [40]:

- G_F is linear and bounded.
- G_F is self-adjoint, $G_F^H = G_F$.
- G_F has $rank(G) = v$.
- G is positive and semi-definite.

The squared Frobenius norm of the Gramian matrix GF is the sum of the squared correlations (SSC) of frame vectors. Therefore, its depreciation is related to the design criterion often employed in CD-NOMA schemes to optimize active user multiplexing and inherent system performance, such as MUSA and PDMA.

However, a more critical result regarding frame incoherence can be obtained by utilizing WB [44]:

$$\mu(F) \geq \sqrt{\frac{b-v}{v(b-1)}} \tag{12}$$

In contrast to the FP, the mutual coherence bound (12) is not always achievable [24]. The only frames that achieve WB are the ETFs. However, ETFs are limited to dimensional pairs (v, b) [45].

An ETF is the closest equivalent of orthonormal bases in terms of incoherence [46], meaning it is a unit-norm tight frame (UNTF) with the maximal spatial angular displacement of frame vector pairs $(f_j \cdot f_i)$ for $j \neq i$ [41]. In other words, ETF achieves WB equality.

A frame is said to be an ETF if it satisfies the following conditions [2]:

1) $\|f_j\|_2 = 1$ for $j = 1, 2, \ldots, b$
2) $\left| (f_j \cdot f_i) \right| = c$ for $j \neq I$ and c is a constant
3) $FF^H = \frac{b}{v} I$

Our previous Steiner Triple System (STS) design [16], as well as our current SOMA design, can be extended into ETFs by modifying the Hadamard matrix [28].

The SOMA(n, k) structure has $\|f_j\|_2 = k$, $\left| (f_j \cdot f_i) \right| = 0$ for users in the same parallel class group and $\left| (f_j \cdot f_i) \right| = 1$ for

users in different parallel classes. These results show that the SOMA design allocated users in the same parallel class as orthogonal and users between other parallel classes with the least coherence. Furthermore, Figure 5. shows an intensity map of the Gramian matrix of SOMA(3,5) and STS(9). These Gramian matrices show that our SOMA designs have similar structures on the rows to STS columns. Following the same method as in [28], SOMA designs can be expanded into ETFs.

Hence, the incidence matrix provided by our SOMA design can increase the efficiency and performance while decreasing the complexity of the modulation process, such as the MPA used in SCMA.

In summary, the measurements of frame theory have provided us with insight into how to design tight, low-coherence, and equal normal frames by the matrix parameter, such as OC from frame redundancy, user fairness from unit-normality, and the angular of column vectors from incoherence. The sparsity and consistent weight with the unique subsets of SOMA provide a unique mapping pattern for users and resources. This matrix has equal fairness and regularity, along with high tightness. Based on these properties, the incidence matrix generated by our SOMA design is also effective when used in other CD-NOMA schemes.

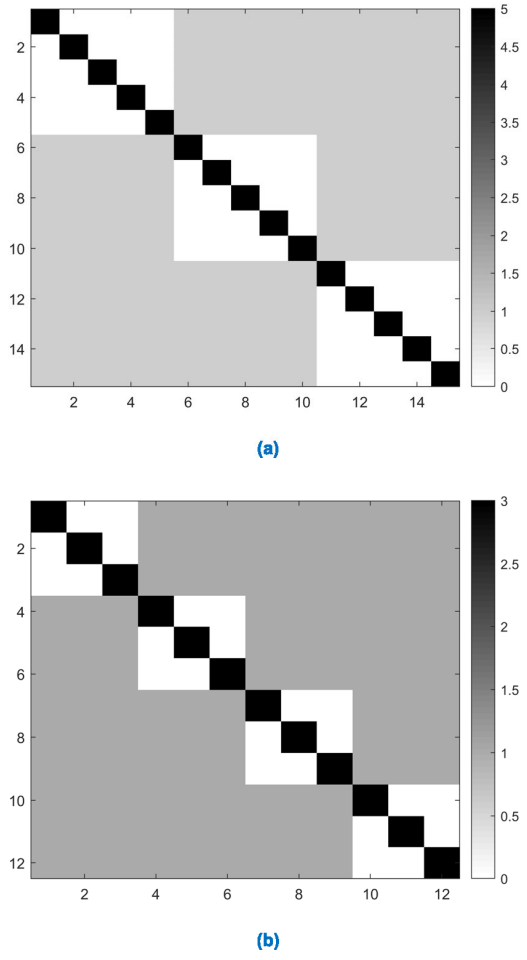


FIGURE 5. (a) Intensity map of Gramian matrix of transpose of SOMA(3,5) (b) Gramian matrix of STS(9).

D. OPTIMIZATION

One of the goals of NOMA is to provide the most services to as many users as possible with the fewest resources under graceful performance. Hence, we can list an objective function by maximizing OC with the constraint of the total power provided by the base station and minimum performance requirements, including the bit/symbol error rate (BER/SER) from [47] and [48], and system throughput, T.

In SOMA designs, SOMA(k, n) can be expressed as $D(v = nk, b = n^2, k = k, r = n)$ where an arbitrary incidence matrix D(v, b, k, r) can be considered as v resources, b users, k resources per user, and r users per resource. We want to increase b and lower v for OC and increase k and lower r for the bit rate and error rate. However, the matrix has a relationship of $k \cdot b = r \cdot v$, so we need to study the tradeoff between these parameters.

Assume that each user transmits m bits per resource block. The variables are represented as follows: P is power and γ is the signal-to-noise ratio (SNR). Therefore, the optimization problem can be mathematically formulated as follows:

$$\max OC = \frac{b}{v} = \frac{n}{k}$$

$$s.t. \sum_{i=1}^{v} P_i \le P_{max} \quad \text{(C1)}$$

$$P_i > 0, \forall i \in \{1, 2, \ldots, v\} \quad \text{(C2)}$$

$$k = k_d + k_m \quad \text{(C3)}$$

$$BER_j \le BER_{max}, \forall j \in \{1, 2, \ldots, b\}$$

$$M = 2^{mr} \quad \text{(C4)}$$

$$BER = \frac{1}{r}\frac{\sqrt{M}-1}{M}\left\{\left(\sqrt{M}-1\right) + 4I_1 - \left(\sqrt{M}-1\right)I_2\right\} \text{[47]}$$

$$I_1 = \left(\frac{1-u}{2}\right)^{k_d} \sum_{m=0}^{k_d-1} \binom{k_d-1+m}{m} \left(\frac{1+u}{2}\right)^m$$

$$u = \sqrt{\frac{\beta\bar{\gamma}_i}{1+\beta\bar{\gamma}_i}}; \beta = \frac{3r}{2\,(M-1)}$$

$$I_2 = \frac{4}{\pi} \sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1} \left(1 + \frac{1}{\beta\bar{\gamma}_i}\right)^{-(n+1)} \times \left\{\sum_{j=0}^{k_d-1} \binom{n+j}{j} (1+\beta\bar{\gamma}_i)^{-j}\right\}$$

$$\bar{\gamma}_i = \alpha^2 \frac{E_b}{N_o}$$

α is a random value from the channel condition

$$T = \sum_{i}^{v} (1 - SER_i) \cdot M - b \cdot k_d \cdot 2^m \geq T_{min} \quad \text{(C5)}$$

$$\text{SER} = 1 - \left(1 - 2 \cdot \left(1 - \frac{1}{\sqrt{\text{M}}}\right) \cdot \text{Q}\left(\sqrt{\frac{3 \cdot \log_2 \text{M}}{\text{M} - 1} \cdot \frac{\text{Eb}}{\text{No}}}\right)\right)^2 \text{for r} = 2 \cdot \text{Z}^+ [48]$$

$$\text{n} \in \textit{prime power}; k \leq n - 1 \quad \text{(C6)}$$

$$v = nk, b = n^2, k = k, r = n \quad \text{(C7)}$$

The size of the OC in the SOMA design is determined by n and k. The value n will affect the error rate because it is equal to the number of users per resource (RE), and it is known that the larger the number of users in an RE, the harder it is to distinguish the multiusers' data. On the other hand, it is well known that the more REs a user has, the better the performance, including the error and data rates. Hence, the higher the value of k, the higher the value of n that can be supported based on the performance improvement. However, to achieve a higher OC, we want to increase n but decrease k. Hence, we need to determine the tradeoff between k and n based on different NOMA schemes, channel state information (CSI), and power and performance constraints.

The constraint in (C1) controls the total power that is not exceeded. The term P_{max} is the total power budget, and P_i denotes the transmitted power of each RE. The constraint in (C2) maintains the power of each RE non-negative. Furthermore, because each user is given k resources that can be freely assigned to carry the same or different data, a combination of k should also be considered. The challenge of making the optimum tradeoff on a given k is how we use the most beneficial resources to the user in terms of diversity gain (performance) and spatial multiplexing gain (data rate). Therefore, k can be written as $k = k_d + k_m$ where k_d represents the resources used for diversity and k_m represents multiplexing in (C3).

The constraint in (C4) guarantees that the maximum error rate of each user will not be passed, and the error rate has a positive correlation with k_d and a negative correlation with n. Constraint in (C5) ensures that the efficiency of system throughput per symbol time is above a certain threshold. Because some REs can be used for diversity, we must minus the repeating bits in calculating the throughput.

Finally, to determine the optimal combination of k_d and k_m, we can deploy the diversity-multiplexing tradeoff (DMT) [49]. DMT is used to describe the relationship between boosting the reliability of reception for a given data rate (providing diversity gain) and boosting the data rate for a given reliability of reception (providing multiplexing or degrees of freedom gain) under a fixed set of REs [50].

By adjusting the number of REs sending the repeating data as diversity and the number of REs sending the sequential data as multiplexing on the set of REs given to each user, we simulated the tradeoff as a DMT problem [51]. Thus, we can draw an optimal tradeoff curve to express decision making based on the conditions below.

A diversity gain d* (g) is achieved at multiplexing gain g if

1) Rate R = r· log SNR
2) Outage probability $\text{P}_{\text{out}}(\text{R}) \approx \text{SNR}^{-\text{d}^*(\text{r})}$

Or

3) $\frac{\log \text{P}_{\text{out}}(\text{r} \cdot \log \text{SNR})}{\log \text{SNR}} = -\text{d}^*(\text{r})$

The curve d* (g) represents the diversity–multiplexing tradeoff of the slow-fading channel. Using the curve d* (g), we can determine the optimum composition for the number of REs in either diversity or multiplexing under different situations.

E. COMPUTING COMPLEXITY

Most NOMA research focuses on computational complexity analysis of the receiver. Therefore, we also compare the complexity of our design using the NOMA scheme proposed in our previous work [17] to other decoders.

The NOMA proposed in [17] only needs to search for the union of each user's data, which will form a new multiuser symbol on the constellation. If the summation of the data of all users equals zero, the maximum vector is required to perform the search. Therefore, we proved that the computational complexity is $O(b^2)$ in [52], where b is the number of users. Hereafter, we compare the decoder complexity in Table 1 with the contents of [10] and [19].

TABLE 1. Decoder complexity table.

Receiver	Complexity
MMSE-SIC	$O(v^3 K)$
MPA	$O\left(vM_p^{d_f}\right)$
SIC-MPA	$O\left(vM_p^{d_f}\right)$
EPA	$O\left(vM_p d_f\right)$
Proposed decoder	$O(b^2)$

where K is the number of layers, M_p is the number of projection points on the constellation. For BPSK, $M_p = 2$, d_f = degree of signal superposition on a resource element for MPA, d_c = degree of freedom allowed in the SIC-MPA.

Second, we can compare the computational complexity of constructing the incidence matrix using the SOMA structure with that of SCMA. Because the SOMA(k, n) design is composed by superimposing k MOLS of order n, the complexity can be written as $O\left(kn^2\right)$ as the complexity of generating a MOLS(n) is $O\left(n^2\right)$. On the other hand, according to [53], the complexity of designing a (v, b) SCMA codebook is $(v\,(v-1)+1)^b$ 1, which can be seen as $O\left(v^{2b}\right)$. Therefore, we know that our matrix design has lower complexity than the codebook design of SCMA.

Finally, we found that the computational complexity of completing a partial Latin Square can be exploited as a security property. It is well known that completing a partial

Latin Square is an NP-complete problem [54]. However, if the partial part is the critical set of Latin squares, the search process becomes a unidirectional path. The critical set is a completable partial Latin Square that carries the minimum information needed to uniquely reconstruct the original Latin Square [55]. The partial Latin square is no longer uniquely completable if any entry in the critical set is missing. Therefore, we can utilize the critical set in the authentication, which prevents the adversary under a security gap, a physical layer security metric, to obtain the incidence matrix directly.

F. ADDITIONAL MODIFICATION

In addition to optimizing the general SOMA design, we can also consider modifications that target the b, k, and r parameters.

Increasing b improves the user support capability. Therefore, the SOMA structure must be expandable to achieve massive user transmissions. We introduced two basic methods for expanding the system.

The first method is to repeat the incidence matrix diagonally and pad the new matrix with 0, as shown in (13). Using this method, the performance, OC, and complexity remained the same as those of the base matrix.

$$\mathrm{D}(30, 50, 5, 3) = \begin{vmatrix} \mathrm{D}(15, 25, 5, 3) & 0 \\ 0 & \mathrm{D}(15, 25, 5, 3) \end{vmatrix} \quad (13)$$

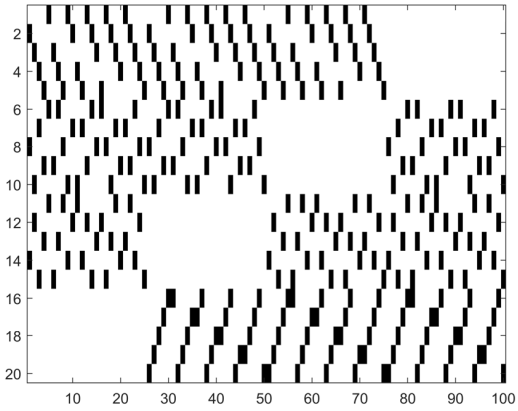
The second method is to combine the SOMA of the same order formed by the permutations of the MOLS side by side. Thus, the system can be rewritten as q-SOMA(k, n), where q is the number of SOMA(k, n) used, k is the number of MOLSs used in each SOMA, t is the total number of MOLS, and n is the value of the order of the MOLS.

As an example, we can construct 4-SOMA(3, 5) as an incidence matrix that supports 100 users by 20 frequencies—D(20,100,15,3)—by adding 4 different SOMA(3, 5) side by side, as shown in Figure 6. where 6a illustrates its structure, and 7b illustrates its sparsity. As shown in Figure 6a., there was a total of $t = 4$ MOLSs. Therefore, we can determine the frequencies by using $t \cdot n = 4 \cdot 5 = 20$. Similarly, the total number of users supported by system 4-SOMA(3, 5) can be determined as $q \cdot n^2 = 4 \cdot 25 = 100$. This example is chosen because it uniformly utilizes all combinations of 3 from the complete set of MOLS(6). Therefore, it represents the incidence matrix with the largest OC we can get from SOMA(3, 5) without breaking the definition of BIBD, which helps it to keep the characteristic of near ETF.

In our previous study [18], we developed an expurgation algorithm for BIBD designs in order to combat constellation expansion. The algorithm decreases r by decreasing k. We further adjusted the algorithm to utilize the resolvable property of SOMA designs by expurgating the RE from parallel classes of different MOLS. Using this method can help maintain the balance of each resource and generate user groups while limiting constellation expansion. An example of expurgating r by 1 in the SOMA(3, 5) design is presented below. We can obtain a system that supports 15 users with

MOLS1	MOLS1	MOLS1	0
MOLS2	MOLS2	0	MOLS2
MOLS3	0	MOLS3	MOLS3
0	MOLS4	MOLS4	MOLS4

(a)



(b)

FIGURE 6. (a) Example 4-SOMA(3, 5) structure. (b) Example 4-SOMA(3, 5) intensity map.

k = 2 and 10 users with k = 3, while maintaining r = 4 after expurgation. User hierarchy can therefore be developed by arranging the user with higher requirements to 10 columns with 3 REs, and the others to 15 columns with 2 REs in the incidence matrix.

MOLS1

3	4	5	1	2
4	5	1	2	3
5	1	2	3	4
1	2	3	4	5

MOLS2

8	10	7	9	6
10	7	9	6	8
6	8	10	7	9
7	9	6	8	10

MOLS3

14	12	15	13	11
15	13	11	14	12
12	15	13	11	14
13	11	14	12	15

(a)

8,14	10,12	7,15	9,13	6,11
3,15	4,13	5,11	1,14	2,12
4,10	5,7	1,9	2,6	3,8
5,6,12	1,8,15	2,10,13	3,7,11	4,9,14
1,7,13	2,9,11	3,6,14	4,8,12	5,10,15

(b)

FIGURE 7. (a) Expurgation of parallel classes by MOLS (b) Expurgated SOMA(3, 5) with 15 users and 2 REs in gray and 10 users and 3 REs in white.

III. NOMA SYSTEM AND APPLICATION

In this section, we discuss an application that utilizes the characteristics of the SOMA design.

A. NOMA SYSTEM

The first application is the NOMA scheme. In our previous work [17], a NOMA downlink (DL) system, shown in Figure 8, was considered. The system can utilize a BIBD incidence matrix D(v, b, k, r), which has a total of b users

and v resources, where $v < b$. In addition, the system distributes the resources in a balanced manner required by most NOMA systems, such that it allocates r users per resource and k resources per user. Although resources can be seen as many things, such as powers, sparse codes, and spaces, we treat each RE as different orthogonal frequency-division multiplexing (OFDM) subcarriers in our NOMA system.

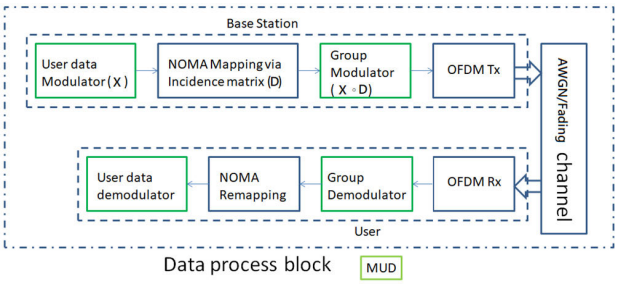


FIGURE 8. System Model.

The multiuser data can be sent by the first union of the information bits m_i of each user per resource to form new complex data and map the complex data to a M-quadrature amplitude modulation (QAM) constellation to create a new complex OFDM symbol.

$$y_j = \bigcup_{i=1}^{r} m_{i,j} \quad where \quad j = 1 \sim v \tag{14}$$

After OFDM demodulation is performed at the receiver, the received symbol with v orthogonal tones can be written as

$$\mathbf{r}_{\mathrm{DL}} = \mathbf{h} \cdot \mathbf{Y} + \mathbf{N} \tag{15}$$

where $\mathbf{h}$ denotes the channel-coefficient matrix. $\mathbf{Y}$ and $\mathbf{N}$ represent the information vector and the noise, respectively. User data are transmitted in k resources; only k out of v resources are decoded per user, thereby decreasing decoding complexity.

Hereafter, we can exploit multiuser detection (MUD), as shown in Figure 9. at the receiver utilizing the incidence matrix structure with parallel classes user allocations without using successive interference cancellation (SIC) or message passing algorithm (MPA). Currently, the major MUD algorithms in NOMA diverge in two ways. The first is serial decoding such as SIC. The second is iterative decoding, such as in MPA. Because the SIC method must wait for the decoding of each user, its latency is higher than that of MPA. However, the calculation of MPA is more complex than that of SIC [56]. Nevertheless, in our design, users only need to know the order of their information bits in the complex OFDM symbol to decode multiuser data. Therefore, our MUD can process the observed data in parallel with low complexity, where no step-to-step serial or iterative processes are involved.

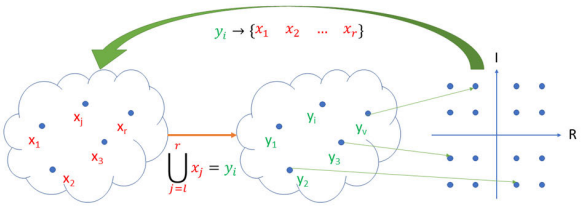


FIGURE 9. The two-stage MUD.

B. BACKWARD COMPATIBILITY AND DYNAMIC ADAPTION

In our NOMA system, signals are transmitted via OFDM subcarriers, and our MUD can decode multiuser data based on their location in the complex OFDM symbol [17]. Thus, we can apply our system design by upgrading the firmware level of many existing Internet of Things (IoT) devices. Making our design as cost-effective as OFDM and QAM technologies and equipment is already highly mature.

The second benefit of using this design is that we can apply the well-known performance analysis to the system owing to our MUD and the preprocessing of the user resource allocation structure of the incidence matrix. Identifying performance criteria is essential for establishing a good system design. Other NOMA designs, such as SCMA, have difficulty in determining their exact BER owing to their multichannel transmissions and complex MUDs.

Generally, QAM performs well in terms of channel capacity to the extent that it is close to Shannon's limit. Therefore, we chose QAM as the default modulation method for both the user and group modulation in this study. We can obtain the well-known BER equation of M-square QAM in AWGN from [57], where M is the size of the QAM, and the equation for the Rayleigh channel from [58] and [59].

Moreover, because our MUD can decode the user's message in parallel without an iterative process, we can choose between sending repeated data on different REs for performance gains or sequential data on different REs for rate gains. We chose the maximal ratio combining (MRC) to support diversity by combining the k_d replicas of user data in the NOMA system because it is an optimal combining method based on the maximum likelihood of the channel [60]. By combining and transmitting multiuser data in OFDM symbols, we can avoid the weaknesses in MRC – poor at handling multiuser interference – thus making MRC a superior candidate as a diversity combiner.

The BER equations of QAM with k-branches of the MRC are also well known and can be found in [47].

However, unlike the conventional MRC procedure, our system model requires that the same user data spread over k branches of REs be demodulated before combination. This swap is necessary, because each received signal contains data from a different set of users. Therefore, they cannot be combined directly, leading to minor performance

degradation. We found that the loss is approximately 2–3 dB in the fading channel and 0–1 dB in the AWGN channel versus the theoretical BER performance using Monte Carlo simulations.

Because we can predict the performance of our NOMA system, we can find the tradeoff between the signal-to-noise ratio (SNR), which is defined as the energy per bit divided by the noise power (Eb/No), and the size of the incidence matrix under different conditions. In future work, we will apply machine learning algorithms to automatically adjust the incidence matrix by monitoring the error rate. Therefore, we can perform resource optimization based on theoretical analysis and targeted performance to achieve dynamic adaptation.

C. USER GROUPING AND HIERARCHY

The SOMA design can be considered a resolvable BIBD. Hence, we can provide neutral user grouping based on the resolvable property of the incidence matrix structure. The resolvable property generates parallel classes, as defined in hypergraph theory and combinatorial designs. The parallel class is also called perfect matching in a hypergraph. A matching in a hypergraph is a collection of disjoint edges, and perfect matching covers the entire vertex set [61]. Therefore, each user group in the matrix will have a balanced distribution for every REs.

Using frame theory, we found that users in the same parallel class were allocated orthogonally to each other in our SOMA design. Consequently, we can assign users who are close in distance to the group of the same parallel class, so they would have minor interference.

We can also extend the concept of parallel class grouping to a user hierarchy. A neutral user hierarchy forms when the expurgation of the incidence matrix does not include every parallel class due to the different k-values in each group leading to differences in performance.

Another type of hierarchy can be formed by changing the modulation of the NOMA system. By combining the concepts of constellation modulation [62] (decreasing the size of the constellation by using multiple constellations to carry information), trellis modulation [63] (increasing the size of the constellation to increase the performance), and the balanced distributed incidence matrix of NOMA, we can provide a user hierarchy based on the performance difference gained by the successive decoding process.

In the next section, we simulate a prototype of constellation trellis modulation that divides QAM constellations by trellis and maps the information to constellation partitions. Because of the diversity of the incidence matrix, unlike the traditional way of utilizing a constellation as an information medium, we do not need to repeat the constellation for the receiver to identify the correct message carried by the constellation. Therefore, the data rate of the group using constellations to receive their message can be increased to be on par with the user using symbols.

D. SECURITY

SOMA is composed of Latin squares, which are used in cryptography owing to their large number of patterns and reconstruction ability. Unlike STS, which has only one resolution pattern, the SOMA structure can provide higher security because of its abundant grouping patterns of incidence matrices.

By utilizing the concept of critical sets, we only need to transmit partial information to the user during authentication to restore the full incidence matrix. The adversary cannot reconstruct the incidence matrix without guessing in the NP-complete problem if there is a security gap [64], which will cause some missing information on the received critical set.

The BER performances of the legal user, Bob, and the eavesdropper, Eve, are used to quantify the secrecy gap, which is easier to analyze than the secrecy capacity. The difference between Bob's SNR (which is required for reliable decoding for a particular service) and Eve's SNR (which is not sufficient to achieve reliable decoding for the same service) reflects Bob's channel-quality advantage over Eve in order to satisfy the practical secrecy notion (i.e., security gap).). The security gap is defined as

$$S_g = SNR_{min}^{Bob} - SNR_{max}^{Eve} \tag{16}$$

This gap between these two SNR levels reflects the channel quality advantage that Bob must possess over Eve to satisfy the practical notion of transmission secrecy.

In addition, all the user bits in the complex OFDM symbol will serve as artificial noise against the adversary as long as they do not know the order of the targeted user inside the symbol or the user grouping per resource from the incidence matrix. Furthermore, we can interleave the order in the complex OFDM symbol generated by the union of user data per resource.

For example, the symbol of $r = 4$ users per RE can be written as $\{U_t\ U_2\ U_3\ U_4\}$, and we assume that the eavesdropper wants to listen to the targeted user, U_t, while all users have L = 3 bits of data to send and send 1 bit per symbol time. As the symbol order is swappable, the actual symbol can be $\{U_t\ U_2\ U_3\ U_4\} \rightarrow \{U_2\ U_3\ U_t\ U_4\} \rightarrow \{U_4\ U_2\ U_t\ U_3\}$. Hence, the eavesdropper can only decode a targeted user's data if he/she guesses it correctly by the chance of $\left(\frac{1}{r}\right)^L$, where L is the length of the information. Therefore, we know that the longer r and L are, the more challenging it is for an adversary to decode the data.

The user assigned to the columns of the incidence matrix is swappable inside each parallel class. Therefore, user data can appear at different locations inside a symbol by mixing with different users at each symbol time. An adversary will not be able to combine the data into a complete piece of information without knowing the hopping sequence of the user, and he/she will not be able to perform diversity without knowing the incidence matrix structure, thus securing our data. This swapping can occur between users in the same parallel class.

For parallel class user column swapping, the eavesdropper must guess another probability of $\left(\frac{1}{n}\right)^L$ to follow the correct targeted user, where n is the order of SOMA(k, n), which is at the same time the number of users in the same parallel class.

IV. SIMULATION RESULTS AND DISCUSSION

The preliminary results are presented in this section. Monte Carlo simulations were performed to evaluate the performance using MATLAB®. Simulations were performed over the AWGN and Rayleigh fading channels.

The simulation parameters are listed in the following table:

TABLE 2. Simulation parameters.

Parameter	Value
User Modulation	BPSK, QPSK
Group Modulation	M-QAM
Channel condition	normalized AWGN normalized flat fading Rayleigh
# of frequencies	4, 8, 12, 15, 20, 32
# of users	6, 16, 25, 48, 64, 100
Matrix Dimensions	4x6, 8x16, 12x16, 12x48, 15x25, 20x25, 32x64, 20x100

A. THEORETICAL PERFORMANCE SIMULATION

The first figure in this section is shown in Figure 10. compared the theoretical performance equations and Monte Carlo simulations of our NOMA system using BPSK in both the AWGN and Rayleigh channels. The results are shown in Figure 10. validate our observations and statements: We can use the well-known analysis equation directly, and the theoretical BER with MRC will degrade slightly such that it needs to be adjusted by 2 dB in Rayleigh fading and 1 dB in AWGN.

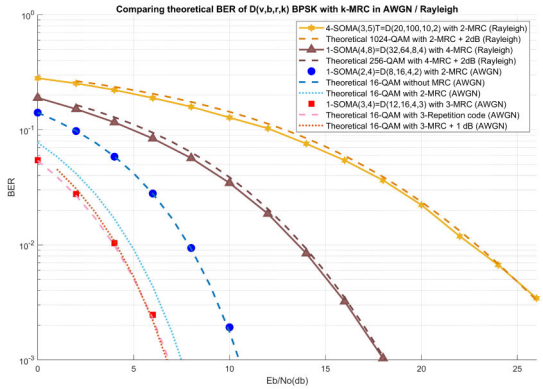


FIGURE 10. Comparing NOMA simulation performance with theoretical BER equation [59] in AWGN / Rayleigh.

B. USER HIERARCHY SIMULATION

Numerous simulations were performed using prototype constellation trellis modulation. The preliminary results of the small incidence matrix SOMA(2,4) are shown in Figure 11. The SOMA array can be expressed as D(v=8,b=16,r=4,k=2), where v is the total resources, b is the total number of users, r is the number of users per resource, and k is the number of resources per user. Utilizing parallel classes, we can categorize users into two groups, where each group has 2 users per resource.

The BER shown in Figure 11, we found that the system performance is similar to that of OMA with 4 bits per user and SOMA(2, 4) without constellation trellis modulation. On the other hand, the higher-priority group performs similarly to OMA, with 2 bits per user. The difference between the high and low user groups was approximately 6 dB.

Overall, we found that the simulations show that user hierarchy can be incorporated into NOMA as an innate system, and that the system's performance, on average, will not be significantly affected.

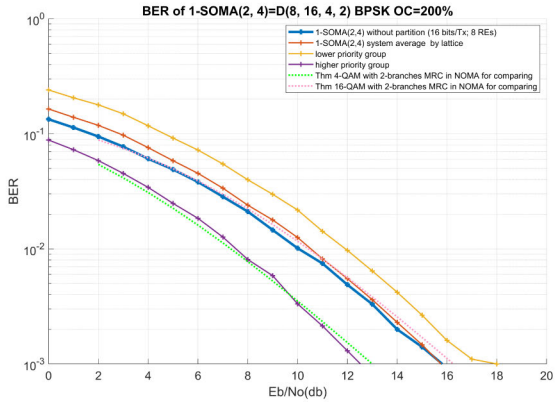


FIGURE 11. Prototype of SOMA(2,4) forming two user hierarchies with constellation trellis modulation.

C. SECURITY

The SOMA design with security was simulated to show the effectiveness of the interleaving symbol order and user allocation. In Figure 12, we found that the eavesdropper cannot

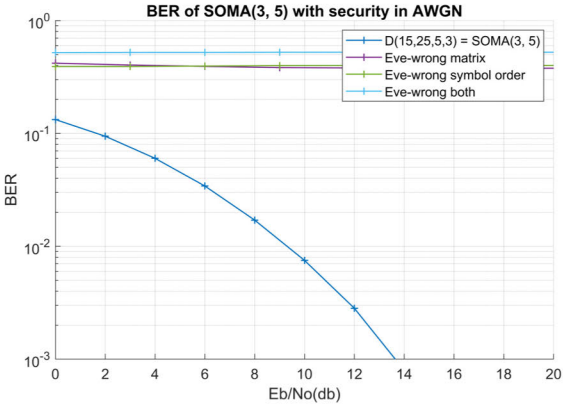


FIGURE 12. Security of SOMA(3,5).

decode the targeted user's data if he missed any hopping sequence on symbol order or user combination.

D. COMPARE THE PERFORMANCE OF THE SAME 4 × 6 MATRIX TO OTHER NOMA SCHEMES

We compared the performance of NOMA utilizing different SOMA sizes with that of other NOMA schemes. Signals marked with -T, such as 3-SOMA(2,4)T, in the results were a system with an incidence matrix expurgated via transversal introduced in Section II-F.

An incidence matrix of D(4,6,3,2) was included as a reference for comparison with SCMA [19], PDMA [10], MUSA [20], and power-imbalance LDS [65], all of which had the same 4 × 6 matrix. The throughput per system is also calculated.

In Figure 13, the results for BPSK or QPSK transmissions per user in AWGN channels for different matrix sizes and OCs are illustrated. Moreover, we compared our SOMA designs to SCMA [18], which has a 2 bits per user transmission rate.

We found that our D(4,6,3,2) system in multiplexing mode outperformed SCMA [19] and LDS [65]. Unfortunately, the authors of [14] did not demonstrate their performance in an AWGN channel. The LDS design is key in [14] for generating larger incidence matrices, such as 7 × 9 and 13 × 15, where both trade the OC for performance. However, comparing the 1/3 rate turbo-coded system, our 3-SOMA(2,4)T design outperforms the LDS 13 × 15 in [15] in terms of both BER and data rate.

Furthermore, we observe the left group of lines in Figure 13. that 3-SOMA(2,4)T (OC=400%) using BPSK performed similarly to SCMA (OC=150%) using QPSK. 3-SOMA(2,4)T (OC=400%) using BPSK can transmit 48 bits at a time using 12 REs, whereas SCMA (OC=150%) using QPSK can transmit 12 bits at a time using 4 REs. It should be noted that even when the number of REs is equalized, 3-SOMA(2,4)T provides better bits per Ts per system. Calculations indicate that our NOMA design can enhance user transmission by 33% over the SCMA design in [19], and enhance user support by 166%. Meanwhile, the simulation results in the middle group of data indicate that constellation expurgation is effective because we were able to find comparable performance between D(4,6,3,2) using QPSK and 4-SOMA(3,5)T using BPSK, which supports 30 and 100 users using 20 REs, respectively. Similarly, the right group depicts a design that uses 3-SOMA(2,4)T with QPSK to send 98 bits per Ts per system by 12 REs and achieve BER $= 10^{-3}$ below $\frac{\text{Eb}}{\text{No}} = 25$dB.

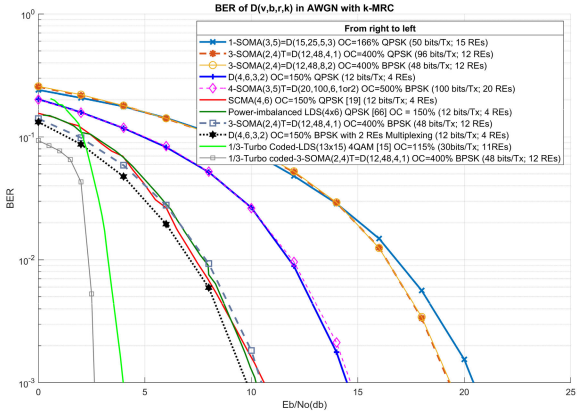


FIGURE 13. BER with MRC vs. Eb/No in AWGN.

Figure 14. illustrates the Rayleigh fading performance of a downlink channel. The following comparison can be observed in the Rayleigh distribution between D(4,6,3,2), SCMA, PDMA, MUSA, and LDS. Our study finds that D(4,6,3,2) outperforms both PDMA and MUSA, while being on par with SCMA. We also present designs that achieve BER $= 10^{-3}$ at medium (25 dB) and high (35 dB) Eb/No, namely, 1-SOMA(3,5) and 3-SOMA(2,4)T, respectively. In contrast to AWGN, the diversity mode of our system plays a vital role in allowing reference design D(4,6,3,2) to outperform other NOMA schemes. Therefore, we are able to provide a design, 1-SOMA(2,4), which increases both the transmission rate and user support by 33% compared with the other four schemes while retaining a comparable SNR.

Furthermore, as shown in Figure 15, we presented the throughput of our system with designs aimed at medium and high SNRs in the Rayleigh channel. We found that, even after scaling to have the same number of REs, our designs still provide better throughput. Based on our simulation results, our reference design, D(4,6,3,2), has a throughput comparable to that of SCMA and PDMA [65]. Meanwhile, 1-SOMA(3,5) and 3-SOMA(2,4)T reach their maximum throughputs at the corresponding Eb/No.

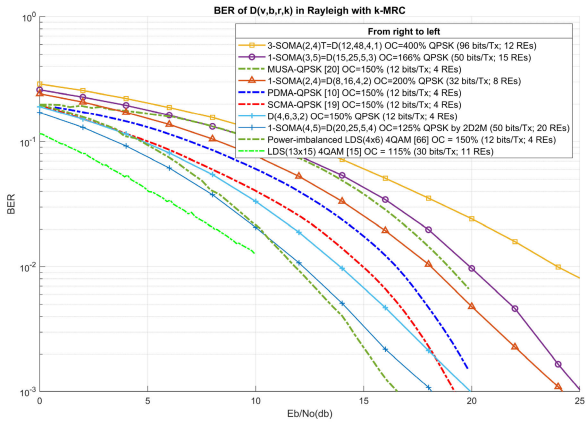


FIGURE 14. BER with MRC vs. Eb/No in Rayleigh.

E. RELATED PAPERS DISCUSSION

While the majority of papers analyzed a low-dimension 4 × 6 matrix, recent papers, such as [14], [15], and [16], have considered some larger incidence matrix systems. Therefore, it is reasonable to compare the results of both studies in terms of performance, OC, latency, and complexity with our results.

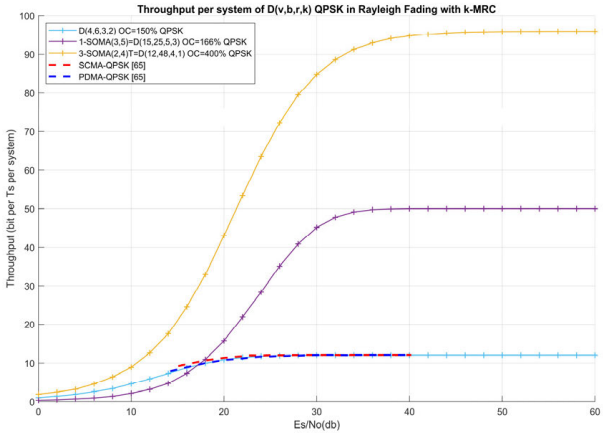


FIGURE 15. Throughput (bit per Ts per system) vs. Es/No in Rayleigh.

However, the performance and complexity of [16] remain the same as those of the 4×6 SCMA. Therefore, this comparison can be skipped as we already compared with 4×6 systems in the previous subsection.

In [15], a power-imbalanced LDS design from [66] was utilized and combined with Singer's theorem [67] to construct a larger incidence matrix, such as 7×9 and 13×15, with an OC of 125% and 115%, respectively. In contrast, the authors of [14] considered maximum distance separable (MDS) codes to form their incidence matrix. An example of a 16×20 resource allocation matrix with OC $= 125\%$ is presented using the MDS method.

Performance-wise, [15] is advantageous not only based on the minimum distance criterion but also on the average Gaussian separability margin, whereas the design in [14] considers finding a larger girth and less dense matrices to improve the performance of the MPA decoder. Both [14] and [15] displayed excellent performance and surpassed many of our designs in terms of Rayleigh channels. Nevertheless, they lower OC in exchange for better performance. The OCs of their designs were all smaller than the most commonly analyzed 4×6 matrix. However, a critical aspect of the NOMA technique is maintaining graceful performance while offering a higher OC that our design can easily support.

To demonstrate this conclusion, we compared our 1-SOMA(4,5) design, where 2 REs out of 4 REs are used for spatial multiplexing, and the remaining 2 REs are used as diversity, as shown in Figure 14. with a 13×15 LDS system [15]. Our 1-SOMA(4,5) design, compared with the LDS system, has a 2 dB loss at BER $= 10^{-2}$ around SNR $= 10$ dB under the same transmission rate per user, but has a higher OC, 125% vs. 115%. This BER was chosen because the authors in [15] did not provide further uncoded performance in Rayleigh after SNR $= 10$ dB. In our view, it is more important to offer a better OC at an acceptable performance than to obtain better results in a near-orthogonal incidence matrix.

In addition, we found that our MUD decoder has better latency than that in [14] because we can decode the message in a single parallel step, whereas their 56×70 design requires repeating 268 MPA cycles.

In terms of complexity, our MUD encoder and decoder design are similar to those of traditional demodulators such as QAM. Consequently, it is also lower than both [14] and [15], in which MPA and probabilistic data association (PDA) multiuser detectors are used.

V. CONCLUSION

A combinatorial structure approach that supports massive user transmission in NOMA was presented in this paper.

In the first step, we explain why we upgraded our previous STS design to the SOMA structure. We briefly introduce SOMA to build an editable matrix that can account for fair resource allocation and confirm its characteristics through frame theory measurements. Examples of SOMA structures are illustrated and calculated using the frame theory regarding frame redundancy, tightness, incoherence, Gramian matrix operations, and unit normality.

Second, we discuss the objective function for matrix optimization under the power, error rate, and data rate constraints. We calculated the computational complexity and showed that our design had lower receiver and matrix design complexity. Several additional methods for adjusting the matrix were also presented.

Hereafter, we discuss the advantages and applications of the proposed design. The MUD used in our NOMA system can produce a bijection mapping between users and constellations, eliminating the requirement for SIC and MPA while solving the subjective mapping problem. Therefore, it provides a theoretical evaluation of system performance, as confirmed by simulations. Security and user grouping utilizing the properties of Latin squares are discussed.

Finally, in the simulation, we demonstrated that we can implement NOMA systems that support more users with relatively low complexity and comparable or better performance. Our simulation results indicate that our system achieves a similar or better performance than other code-domain NOMA systems. The system can achieve a higher overload by increasing the combination and order of SOMAs as well as solving the problem of constellation expansion by editing SOMAs with expurgation.

REFERENCES

[1] S. Datta and J. Oldroyd, "Low coherence unit norm tight frames," *Linear Multilinear Algebra*, vol. 67, no. 6, pp. 1174–1189, Jun. 2019, doi: 10.1080/03081087.2018.1450348.

[2] M. A. Sustik, J. A. Tropp, I. S. Dhillon, and R. W. Heath Jr., "On the existence of equiangular tight frames," *Linear Algebra Appl.*, vol. 426, nos. 2–3, pp. 619–635, Oct. 2007, doi: 10.1016/j.laa.2007.05.043.

[3] W. U. Bajwa and A. Pezeshki, "Finite frames for sparse signal processing," in *Finite Frames*. New York, NY, USA: Springer, 2013, pp. 303–335, doi: 10.1007/978-0-8176-8373-3_9.

[4] A. Kapur and M. K. Varanasi, "Multiuser detection for overloaded CDMA systems," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1728–1742, Jul. 2003, doi: 10.1109/TIT.2003.813562.

[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, Jul. 2017, doi: 10.1109/jsac.2017.2725519.

[6] H. D. Schotten and H. Hadinejad-Mahram, "Analysis of a CDMA downlink with non-orthogonal spreading sequences for fading channels," in *Proc. IEEE 51st Veh. Technol. Conf. (VTC-Spring)*, May 2000, pp. 1782–1786, doi: 10.1109/vetecs.2000.851579.

[7] K. Deka and S. Sharma, "Hybrid NOMA for future radio access: Design, potentials and limitations," *Wireless Pers. Commun.*, vol. 123, no. 4, pp. 3755–3770, Apr. 2022, doi: 10.1007/s11277-021-09312-3.

[8] M. M. Sahin and H. Arslan, "Waveform-domain NOMA: The future of multiple access," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Jun. 2020, pp. 1–6, doi: 10.1109/iccworkshops49005.2020.9145077.

[9] M. T. P. Le, G. C. Ferrante, G. Caso, L. De Nardis, and M. Di Benedetto, "On information-theoretic limits of code-domain NOMA for 5G," *IET Commun.*, vol. 12, no. 15, pp. 1864–1871, Sep. 2018, doi: 10.1049/iet-com.2018.5241.

[10] 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, Apr. 2017, doi: 10.1109/tvt.2016.2596438.

[11] L. Zhu, J. Zhang, Z. Xiao, X. Cao, and D. O. Wu, "Optimal user pairing for downlink non-orthogonal multiple access (NOMA)," *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 328–331, Apr. 2019, doi: 10.1109/lwc.2018.2853741.

[12] C. B. Mwakwata, O. Elgarhy, M. M. Alam, Y. Le Moullec, S. Parand, K. Trichias, and K. Ramantas, "Cooperative scheduler to enhance massive connectivity in 5G and beyond by minimizing interference in OMA and NOMA," *IEEE Syst. J.*, vol. 16, no. 3, pp. 5044–5055, Sep. 2022, doi: 10.1109/jsyst.2021.3114338.

[13] S. Furino, Y. Miao, and Y. Jianxing, *Frames and Resolvable Designs: Uses, Constructions, and Existence*. Boca Raton, FL, USA: CRC Press, 1996.

[14] B. F. da Silva, D. Silva, B. F. Uchoa-Filho, and D. L. Ruyet, "A multi-stage method for SCMA codebook design based on MDS codes," *IEEE Wireless Commun. Lett.*, vol. 8, no. 6, pp. 1524–1527, Dec. 2019, doi: 10.1109/lwc.2019.2925801.

[15] G. Millar, M. Kulhandjian, A. Alaca, S. Alaca, C. D'Amours, and H. Yanikomeroglu, "Low-density spreading design based on an algebraic scheme for NOMA systems," *IEEE Wireless Commun. Lett.*, vol. 11, no. 4, pp. 698–702, Apr. 2022, doi: 10.1109/lwc.2022.3140223.

[16] L. Li, Z. Ma, P. Z. Fan, and L. Hanzo, "High-dimensional codebook design for the SCMA down link," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 10118–10122, Oct. 2018, doi: 10.1109/tvt.2018.2863552.

[17] Y. Wu, E. Attang, and G. E. Atkin, "Low complexity NOMA system with combined constellations," *IEEE Wireless Commun. Lett.*, vol. 8, no. 5, pp. 1308–1311, Oct. 2019, doi: 10.1109/lwc.2019.2897795.

[18] X. Hao and Y. Wu, "Low density code design for downlink NOMA system," *IEICE Commun. Exp.*, vol. 10, no. 3, pp. 124–130, Mar. 2021, doi: 10.1587/comex.2020xbl0161.

[19] H. Nikopour and H. Baligh, "Sparse code multiple access," in *Proc. IEEE 24th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Sep. 2013, pp. 332–336, doi: 10.1109/pimrc.2013.6666156.

[20] Z. Yuan, G. Yu, W. Li, Y. Yuan, X. Wang, and J. Xu, "Multi-user shared access for Internet of Things," in *Proc. IEEE 83rd Veh. Technol. Conf. (VTC Spring)*, May 2016, pp. 1–5, doi: 10.1109/vtcspring.2016.7504361.

[21] D. Stinson, *Combinatorial Designs*. New York, NY, USA: Springer, 2007.

[22] W. D. Wallis, *Introduction to Combinatorial Designs*, 2nd ed. Boca Raton, FL, USA: CRC Press, 2007.

[23] P. G. Casazza and G. Kutyniok, *Finite Frames*. New York, NY, USA: Springer, 2012.

[24] T. Strohmer, "A note on equiangular tight frames," *Linear Algebra Appl.*, vol. 429, no. 1, pp. 326–330, Jul. 2008, doi: 10.1016/j.laa.2008.02.030.

[25] M. N. Iranzad, "Applications of balanced incomplete block designs to communication systems," Ph.D. dissertation, Univ. Virginia, Charlottesville, VA, USA, 2013, doi: 10.18130/v35229.

[26] C. J. Colbourn, *CRC Handbook of Combinatorial Designs*. Boca Raton, FL, USA: CRC Press, 2010.

[27] A. Ferber and M. Kwan, "Almost all Steiner triple systems are almost resolvable," *Forum Math., Sigma*, vol. 8, p. e39, Nov. 2020, doi: 10.1017/fms.2020.29.

[28] M. Fickus, D. G. Mixon, and J. C. Tremain, "Steiner equiangular tight frames," *Linear Algebra Appl.*, vol. 436, no. 5, pp. 1014–1027, 2012, doi: 10.1016/j.laa.2011.06.027.

[29] A. Thompson and R. Calderbank, "Sparse near-equiangular tight frames with applications in full duplex wireless communication," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Nov. 2017, pp. 868–872, doi: 10.1109/GlobalSIP.2017.8309084.

[30] N. C. K. Phillips and W. D. Wallis, "All solutions to a tournament problem," in *Congressus Numerantium*, vol. 114. Canada: Utilitas Mathematica, 1996, pp. 193–196.

[31] L. H. Soicher, "On the structure and classification of SOMAs: Generalizations of mutually orthogonal Latin squares," *Electron. J. Combinatorics*, vol. 6, no. 1, p. R32, Jul. 1999, doi: 10.37236/1464.

[32] O. Grošek and M. Sýs, "Isotopy of Latin squares in cryptography," *Tatra Mountains Math. Publications*, vol. 45, no. 1, pp. 27–36, Dec. 2010, doi: 10.2478/v10127-010-0003-z.

[33] S. Stein, "Transversals of Latin squares and their generalizations," *Pacific J. Math.*, vol. 59, no. 2, pp. 567–575, Aug. 1975, doi: 10.2140/pjm.1975.59.567.

[34] J. Shen, T. Zhou, X. Liu, and Y.-C. Chang, "A novel Latin-square-based secret sharing for M2M communications," *IEEE Trans. Ind. Informat.*, vol. 14, no. 8, pp. 3659–3668, Aug. 2018, doi: 10.1109/tii.2018.2810840.

[35] B. S. Boob and H. L. Agrawal, "A note on the construction of mutually orthogonal Latin squares," *Biometrics*, vol. 32, no. 1, p. 191, Mar. 1976, doi: 10.2307/2529349.

[36] G. Song, K. Cai, Y. Chi, J. Guo, and J. Cheng, "Super-sparse on-off division multiple access: Replacing repetition with idling," *IEEE Trans. Commun.*, vol. 68, no. 4, pp. 2251–2263, Apr. 2020, doi: 10.1109/tcomm.2020.2965522.

[37] P. G. Casazza, G. Kutyniok, and F. Philipp, "Introduction to finite frame theory," in *Finite Frames*. New York, NY, USA: Springer, 2013, pp. 1–53, doi: 10.1007/978-0-8176-8373-3_1.

[38] Y. Yuan and C. Yan, "NOMA study in 3GPP for 5G," in *Proc. IEEE 10th Int. Symp. Turbo Codes Iterative Inf. Process. (ISTC)*, Dec. 2018, pp. 1–5, doi: 10.1109/istc.2018.8625325.

[39] O. Christensen, *An Introduction to Frames and Riesz Bases*. Cham, Switzerland: Springer, 2016.

[40] R.-A. Stoica, "Frame-theoretic designs for future wireless communications," Ph.D. thesis, Dept. Comput. Sci. Elect. Eng., Jacobs Univ. Bremen, Bremen, Germany, Oct. 2019. [Online]. Available: http://nbn-resolving.org/urn:nbn:de:gbv:579-opus-1008910

[41] S. Waldron, *An Introduction to Finite Tight Frames*. Basel, Switzerland: Birkhäuser, 2018.

[42] J. Kovacevic and A. Chebira, *An Introduction to Frames*. Delft, The Netherlands: Now Publishers, 2008.

[43] J. J. Benedetto and M. Fickus, "Finite normalized tight frames," *Adv. Comput. Math.*, vol. 18, nos. 2–4, pp. 357–385, 2003, doi: 10.1023/a:1021323312367.

[44] T. Strohmer and R. W. Heath Jr., "Grassmannian frames with applications to coding and communication," *Appl. Comput. Harmon. Anal.*, vol. 14, no. 3, pp. 257–275, May 2003, doi: 10.1016/s1063-5203(03)00023-x.

[45] M. Fickus, J. Jasper, E. J. King, and D. G. Mixon, "Equiangular tight frames that contain regular simplices," *Linear Algebra Appl.*, vol. 555, pp. 98–138, Oct. 2018, doi: 10.1016/j.laa.2018.06.004.

[46] M. Haikin, M. Gavish, D. G. Mixon, and R. Zamir, "Asymptotic frame theory for analog coding," *Found. Trends Commun. Inf. Theory*, vol. 18, no. 4, pp. 526–645, 2021, doi: 10.1561/0100000125.

[47] C.-J. Kim, Y.-S. Kim, and C.-Y. Jung, "BER analysis of QAM with MRC space diversity in Rayleigh fading channel," in *Proc. 6th Int. Symp. Pers., Indoor Mobile Radio Commun.*, 1995, pp. 482–485, doi: 10.1109/pimrc.1995.480915.

[48] J. G. Proakis and M. Salehi, *Communication Systems Engineering*. Upper Saddle River, NJ, USA: Prentice-Hall, 2001.

[49] L. Zheng and D. N. C. Tse, "Diversity and multiplexing: A fundamental tradeoff in multiple-antenna channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 5, pp. 1073–1096, May 2003, doi: 10.1109/tit.2003.810646.

[50] X. Li, Y. Zheng, J. Zhang, S. Dang, and A. Nallanathan, "Finite SNR diversity-multiplexing trade-off in NOMA-assisted TW-ABCom systems," Henan Polytech. Univ., Jiaozuo, China, Tech. Rep., 2022, doi: 10.13140/RG.2.2.28074.75201.

[51] D. N. C. Tse, P. Viswanath, and L. Zheng, "Diversity-multiplexing tradeoff in multiple-access channels," *IEEE Trans. Inf. Theory*, vol. 50, no. 9, pp. 1859–1874, Sep. 2004, doi: 10.1109/tit.2004.833347.

[52] E. Attang, "Multiuser communications for network coding and noma," Ph.D. dissertations, Illinois Inst. Technol., Chicago, IL, USA, 2018. [Online]. Available: https://ezproxy.gl.iit.edu/login?url=https://www.proquest.com/dissertations-theses/multiuser-communications-network-coding-noma/docview/2070378071/se-2

[53] S. Chaturvedi, D. N. Anwar, V. A. Bohara, A. Srivastava, and Z. Liu, "Low-complexity codebook design for SCMA-based visible light communication," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 106–118, 2022, doi: 10.1109/ojcoms.2022.3141800.

[54] C. J. Colbourn, "The complexity of completing partial Latin squares," *Discrete Appl. Math.*, vol. 8, no. 1, pp. 25–30, Apr. 1984, doi: 10.1016/0166-218x(84)90075-1.

[55] J. Cooper, D. Donovan, and J. Seberry. (1991). *Latin Squares and Critical Sets of Minimal Size*. [Online]. Available: https://ro.uow.edu.au/infopapers/1056

[56] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. Hanzo, "A survey of non-orthogonal multiple access for 5G," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2294–2323, 3rd Quart., 2018, doi: 10.1109/comst.2018.2835558.

[57] L.-L. Yang and L. Hanzo, "A recursive algorithm for the error probability evaluation of M-QAM," *IEEE Commun. Lett.*, vol. 4, no. 10, pp. 304–306, Oct. 2000, doi: 10.1109/4234.880816.

[58] M. K. Simon and M.-S. Alouini, *Digital Communication Over Fading Channels*. Hoboken, NJ, USA: Wiley, 2005.

[59] E. K. Hall and S. G. Wilson, "Design and analysis of turbo codes on Rayleigh fading channels," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 2, pp. 160–174, Feb. 1998, doi: 10.1109/49.661105.

[60] M. S. Patterh, T. S. Kamal, and B. S. Sohi, "BER performance of MQAM with L-branch MRC diversity reception over correlated Nakagami-m fading channels," *Wireless Commun. Mobile Comput.*, vol. 3, no. 3, pp. 397–406, May 2003, doi: 10.1002/wcm.93.

[61] J. H. Kim, "Perfect matchings in random uniform hypergraphs," *Random Struct. Algorithms*, vol. 23, no. 2, pp. 111–132, Sep. 2003, doi: 10.1002/rsa.10093.

[62] S. S. Dash, F. Pythoud, D. Hillerkuss, B. Baeuerle, A. Josten, P. Leuchtmann, and J. Leuthold, "Constellation modulation—An approach to increase spectral efficiency," *Opt. Exp.*, vol. 25, no. 14, pp. 16310–16331, 2017, doi: 10.1364/OE.25.016310.

[63] L.-F. Wei, "Trellis-coded modulation with multidimensional constellations," *IEEE Trans. Inf. Theory*, vol. IT-33, no. 4, pp. 483–501, Jul. 1987, doi: 10.1109/tit.1987.1057329.

[64] E. Güvenkaya, J. M. Hamamreh, and H. Arslan, "On physical-layer concepts and metrics in secure signal transmission," *Phys. Commun.*, vol. 25, pp. 14–25, Dec. 2017, doi: 10.1016/j.phycom.2017.08.011.

[65] R.-A. Stoica, G. T. F. de Abreu, T. Hara, and K. Ishibashi, "Massively concurrent non-orthogonal multiple access for 5G networks and beyond," *IEEE Access*, vol. 7, pp. 82080–82100, 2019, doi: 10.1109/access.2019.2923646.

[66] Z. Liu, P. Xiao, and Z. Mheich, "Power-imbalanced low-density signatures (LDS) from Eisenstein numbers," in *Proc. IEEE VTS Asia Pacific Wireless Commun. Symp. (APWCS)*, Aug. 2019, pp. 1–5, doi: 10.1109/vts-apwcs.2019.8851632.

[67] J. Singer, "A theorem in finite projective geometry and some applications to number theory," *Trans. Amer. Math. Soc.*, vol. 43, no. 3, pp. 377–385, May 1938, doi: 10.1090/s0002-9947-1938-1501951-4.

ELI HWANG received the B.Eng. degree in electrical engineering from National Tsing Hua University, in 2013, and the M.Sc. degree in telecommunication and network engineering from Southern Methodist University, in 2015. He is currently pursuing the Ph.D. degree in electrical engineering with a focus on communications with the Illinois Institute of Technology. His current research interests include communication theory, information theory, physical layer security, combinatorial design, and modern modulation. He also received the Graduate Certificate in Communication Systems from the Illinois Institute of Technology, in 2019.

XING HAO (Member, IEEE) received the M.S. degree from the Illinois Institute of Technology, Chicago, IL, USA, in 2018, where he is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering. His research interests include non-orthogonal multiple access, modulation techniques, error detection and correction techniques, signal processing, RF energy harvesting, MIMO, and optimization.

GUILLERMO E. ATKIN (Life Senior Member, IEEE) received the B.S. degree in civil electronic engineering from University Federico Santa Maria, in 1974, and the Ph.D. degree in electrical engineering from the University of Waterloo, ON, Canada, in 1986.

From 1974 to 1981, he was a full-time Lecturer with the Electrical Engineering Department, University Federico Santa Maria, and a Consultant at Telecommunications Industry. From 1982 to 1986, he worked as a Research Assistant at the Natural Sciences and Engineering Research Council of Canada. He is currently the Director of the Digital Communication Systems Laboratory of Illinois Institute of Technology. This laboratory is an advanced research facility equipped with personal computers and associated peripherals. The laboratory is also intended to provide a complete interactive environment for simulation-based analysis and design of communication systems. He has published more than 100 technical papers and reports in the communications area. His current interests include bandwidth efficient coding and modulation techniques, wireless communication systems, and applications of information theory concepts to physical layer security in communications networks.