Sparse Signatures with Forward Error Correction Coding for Non-Orthogonal Massive Access

Johannes Dommel, Zoran Utkovski, Petar Popovski and Sławomir Stańczak

Abstract

Wireless communication systems providing massive access for Internet-of-Things (IoT) applications are of growing importance, especially in connection with 5G networks and beyond. Typical massive access scenarios are studied on the basis of a multiple access channel in which a randomly chosen subset of terminals (users) transmit short messages to a common receiver. The transmission of short messages leads to finite blocklength effects and therefore calls for approaches that jointly address the problem of multi-user interference and noise reduction. In this paper, we propose a coding and transmission scheme that combines sparse signature design and finite-size forward error correction (FEC) coding for non-orthogonal massive access that is suitable for low-complexity receiver processing. Our sparse signature construction method relies on the concept of Euler squares and we use graph theory tools to map a sparse allocation of users to resources in a non-orthogonal access model. We evaluate the system performance using an iterative receiver implementation with extrinsic information exchange between a multi-user detector (MUD) based on message passing algorithm (MPA) and user-specific FEC. The proposed construction can be explicitly characterized for a large number of combinations of system parameters, providing a framework for both, non-orthogonal scheduled - and grant-free (massive) random access.

Index Terms

Internet-of-Things (IoT), non-orthogonal massive access, sparse signature design, message passing algorithm, finite blocklength.

J. Dommel (corresponding author, email: johannes.dommel@hhi.fraunhofer.de), Z. Utkovski and S. Stańczak are with the Department of Wireless Communications and Networks, Fraunhofer Heinrich-Hertz-Institute, Berlin, Germany. P. Popovski is with the Department of Electronic Systems, Aalborg University, Aalborg, Denmark.
Fig. 1: Operational regimes of non-orthogonal multiple access protocols and techniques categorized by the asymptotic scaling of resources $N$ with the total number of supported devices $K$, with $\alpha, \beta > 1$ and $0 < \delta < 1$.

I. Introduction

The growing interest in Internet-of-Things (IoT) applications has put massive machine type communications (mMTC) at the focus of the wireless communications research for 5G networks and beyond. mMTC services are characterized by the presence of a potentially massive number of terminals that transmit short packets in a sporadic fashion. Potential mMTC applications are in various domains, ranging from industry and smart home to environmental sensing or logistics. In these scenarios, of interest is the operational regime where the users share some wireless resources in a non-orthogonal fashion. Indeed, in massive connectivity scenarios with short messages the traditional orthogonal schemes are expected to be highly inefficient, mainly due to an inevitable signaling overhead that may significantly outweigh the benefits of orthogonal access. Therefore, the design of suitable non-orthogonal access strategies plays a central role in facilitating efficient and effective utilization of scarce wireless resources. While non-orthogonal access has recently attracted considerable attention by both researchers and practitioners (see e.g. [1]), a unifying framework that structures and classifies the plethora of different approaches investigated in this context is missing. Perhaps the only possible unifying point is in the definition that non-orthogonal access assumes that the same communication resources are simultaneously shared by multiple users, as illustrated in Fig. [1]
includes the scenario where \( K \) users may be simultaneously scheduled or configured on \( N > K \) resources (Regime I), where asymptotically the blocklength \( N \) is sent to infinity before the number of users \( K \) \([2], [3]\). This model may fall within the framework of non-orthogonal access, as the multiple users can be scheduled or configured to share the same physical resource blocks (PRBs). Most of the communication schemes targeting enhanced mobile broadband (eMBB) transmissions, investigated under the framework of non-orthogonal multiple access (NOMA) fall into this category (see, e.g., \[1\] for an overview). In ALOHA models \[4\] for random access, a packet is the smallest, atomic unit of information. The focus in this model lies on the fact that the users sending packets are uncoordinated and any collision is destructive, which means that non-orthogonality is undesirable. Extending over the multiple-access channel (MAC) model, random user activation has been integrated in information-theoretic models by way of partially active users (“T-out-of-N MAC”) \([5], [6]\). A step towards bridging the gap between the two models is the one used in coded random access with successive interference cancellation (SIC) \([7]\), where a randomly active user does not send her packet only once, as in ALOHA, but she transmits several replicas. In a more advanced version of coded random access, instead of sending replicas of the same packet, the sender sends packets that are related through a forward error correction (FEC) code \([7]\). However, for massive IoT, the packets are usually short and the control information, such as the user address, has a size that is comparable to the data size. This was addressed in the information-theoretic model of many-access channel (MnAC) from \([8]\), which accommodates random activation as it allows each transmitter to be active with a certain probability in each block. This solves the problem of encoding control information in an asymptotic regime by allowing the number of users \( K \) to increase proportionally to the blocklength \( N \) (Regime II). The problem of transmitting the user address within a finite (short) packet when \( K \to \infty \) was addressed using a completely different approach in \([9]\). This work gave rise to unsourced random access (U-RA), where all users use the same codebook and thus no user can send a uniquely identifiable address. The goal of the decoder is to output a list that should contain the different messages that were transmitted by the active users. Under this framework, a collision is the event where multiple users transmit the same codewords from the joint codebook. As a consequence, the total number of users \( K \) can be left out of the model, i.e. it can increase without bound with the blocklength \( N \) (Regime III).
A. Related Work

In the context of non-orthogonal access (both scheduled and random, i.e. grant-free), two important aspects arise: (i) the manner in which the information-carrying messages of the individual users are mapped to the shared resources; and (ii) the resulting implications on the receiver side processing, including decoding performance and receiver complexity. The case of random access is particularly challenging for both aspects, as the mapping to the shared resources should correspond to the activation statistics of the users, while the decoding process is complicated by the fact that the set or even the sheer number of transmitting nodes is unknown.

Some of these aspects have been studied in the context of signature design in NOMA, where users are multiplexed on PRBs by applying signatures as patterns for accessing the shared resources, typically concatenated with low-rate error-correcting codes. Low-density code-domain (LDCD) - NOMA is a prominent sub-category of signature-based multiplexing, where low-density signatures (LDS), i.e. sparse spreading codes, are employed for linearly modulating each user’s symbols over shared physical resources [10], [11]. Significant receiver complexity reduction can be achieved by utilizing message passing algorithms (MPAs), which enable user separation even when the received powers are comparable (as opposed to "power-domain NOMA" [12] approaches). For instance, sparse coded multiple access (SCMA) further optimizes the low-density signatures to achieve shaping and coding gains by using multidimensional constellations [11]. These aspects have also been studied in the literature on coded random access with the note that in that case coding is usually performed at the level of packets [13], rather than on symbol level. In the following we will refer to these approaches under the joint term of sparse non-orthogonal access, where the mapping between users and resources can be either regular, where each user occupies a fixed number of resources, and each resource is used by a fixed number of users; or irregular, where the respective numbers are random, and only fixed on average. The optimal spectral efficiency of irregular constructions was shown in [14] to reside below the well-known spectral efficiency of dense random-spreading (RS) [15]. Recently it was shown that, in certain operational regimes, low-density spreading signatures for NOMA based on regular-sparse constructions outperform not only irregular-sparse, but also dense constructions, both for Gaussian channels [16], and for block-fading channels [17]. This observation has

\[^1\]The result stems from the random nature of the user-resource mapping, due to which some users may end up without any designated resources, while some resources may be left unused.
important practical implications, as sparse mappings allow for feasible near-optimal (MPA-based) multiuser receiver implementation where the belief propagation (BP) benefits from the sparsity of the corresponding factor graph. This is in contrast to dense spreading sequences [15], where the complexity of the optimum receiver may be prohibitive for large systems.

The advantage of using sparse mappings in combination with MPA-based processing can be assessed rigorously in the large system limit (LSL), when the number of users goes to infinity while the ratio of users per dimension (overload) is kept constant. Note that in the context of Fig. 1, this approach falls under the category depicted in the middle section (Regime II). It can be shown in the LSL that BP effectively produces the a-posteriori probability of the input given the observed channel output, under the condition that the iterative equation describing BP has a unique fixed point. With this, BP yields an equivalent scalar Gaussian channel from the viewpoint of each user [18], whereby the collective impact of interfering users results in an effective degradation of the signal-to-noise ratio (SNR) of the desired user, also known as multiuser efficiency.

In the context of massive random access, recently proposed protocols based on the concept of U-RA [9] have attracted considerable attention. Under this framework, the users employ the same codebook and collisions are interpreted as the event where multiple users transmit the same codewords. In this scenario, the problem of user identification is separated from the actual data transmission, and the decoder only declares which messages were transmitted, without associating the messages to the users that transmitted them. A coding scheme for the unsourced Gaussian random access channel was proposed in [19]. According to this approach, the available channel uses are split into sub-blocks of equal length, and each active user randomly chooses only one of these sub-blocks, over which it transmits. All users encode their messages using the same codebook obtained by concatenation of two codes: (i) an inner binary linear code, whose goal is to enable the receiver to decode the modulo-2 sum of all codewords transmitted within the same sub-block and (ii) an outer code whose goal is to enable the receiver to recover the individual messages that participated in the modulo-2 sum. An alternative, state-of-the-art scheme for the U-RA channel was proposed in [20] based on sparse regression codes (SPARCs). Accordingly, each user encodes its message into a sparse binary vector, which is then mapped to the transmitted signal vector by performing linear mapping using a shared dictionary which is divided into sections. An outer code is used to assign the symbols to individual messages.
B. Contribution

In contrast to the LSL, the analysis of the non-asymptotic performance of finite-length sparse designs needs to consider the following: (i) the posterior obtained by BP may not converge for a finite-size system and (ii) even if BP converges to a fixed point solution, the decoupling effect might not hold in general when the fixed point solution is not unique, i.e. it might not be possible to decompose the joint decoding into multiple single-user decoding problems (this is typically the case when the system load exceeds a critical threshold value). Motivated by these observations, we propose a coding scheme for non-orthogonal access that combines the principle of sparse access (via sparse signature design), and FEC in the finite blocklength regime.

The proposed code construction results in a regular layer / user resource allocation based on Euler squares, each layer occupying a fixed number of resources and each resource being used by a fixed number of layers. In addition, the construction is flexible in the sense that it can be described explicitly for a large number of combinations of system parameters, i.e. number of layers / users, number of resource elements, number of resources occupied by each layer, number of layers sharing the same resource. We exploit the properties of the sparse construction and propose an iterative receiver structure with quasi-optimal performance with moderate complexity, which is based on belief propagation, where decoding is performed in a turbo-like fashion by exchanging extrinsic information between a multi-user-detection module employing message passing (MPA module), and a FEC module operating on the level of individual users.

We can position our work as one that works with baseband symbols and uses them to transmit both, user identification and data. It can deal with uncoordinated users, while the use of FEC allows reaping gains that stem from the information-theoretic models. A central statement of this work is that in the context of non-orthogonal access with short messages, the finite size coding mandates that multi-user separation and FEC must be considered jointly. Against this background, we evaluate numerically the joint effects of sparse signature design and FEC and characterize the trade-off between system parameters such as sparsity, system load and channel coding rate. Further, we compare the proposed scheme numerically against the state-of-the-art, illustrating the potential of the proposed approach as a viable solution for massive non-orthogonal access, both for scheduled (i.e. grant based) and grant-free transmissions.

The sparse signature design is based on the concept of Euler squares [21], which is motivated by recent works on deterministic binary matrices for compressed sensing [22] and on error-
correcting codes based on *partial geometries* [23]. For analogous discussions on the advantages of general combinatorial code designs for grant-free access in ultra-reliable low-latency communication (URLLC), see, e.g., [24]. A worked-out example of a sparse design is provided in the sequel. We note that, with appropriate parameterization, the coding scheme can be applied both to scheduled and grant-free random access transmissions.

**Example:** To illustrate the approach for the sparse signature design, consider the following example where 9 users share 6 resources, as prescribed by the sparse matrix

\[
F = \begin{pmatrix}
1 & 0 & 0 & | & 0 & 0 & 1 & | & 0 & 1 & 0 \\
0 & 1 & 0 & | & 1 & 0 & 0 & | & 0 & 0 & 1 \\
0 & 0 & 1 & | & 0 & 1 & 0 & | & 1 & 0 & 0 \\
- & - & - & | & - & - & - & | & - & - & - \\
1 & 0 & 0 & | & 0 & 1 & 0 & | & 0 & 0 & 1 \\
0 & 1 & 0 & | & 0 & 0 & 1 & | & 1 & 0 & 0 \\
0 & 0 & 1 & | & 1 & 0 & 0 & | & 0 & 1 & 0 \\
\end{pmatrix}, \tag{1}
\]

where each column of \( F \) corresponds to the pattern by which the user accesses the resources. We note that the matrix \( F \) is the incidence matrix of a bipartite graph (illustrated in Fig. 2a for the above example) that prescribes the message passing procedure at the receiver. In the grant-free scenario with random activation, the message passing procedure is performed on the "pruned" graph where the variable nodes and edges associated with the *inactive* users are removed (as illustrated in Fig. 2b). The specific structure of this construction (which will be discussed in more details in Section III) guarantees that the individual segments of the users only collide at
most on one segment, whereas the total number of devices per segment (as well as the number of segments per device) are fixed.

C. Outline

In Section II, we introduce the system model for sparse non-orthogonal access and describe the building blocks of the applied coding scheme. In addition, we discuss the specifics of the approach, as well as some related aspects from the literature on unsourced random access. In Section III, we provide the theoretical background on graph theory, especially with focus on bi-regular graphs and their representation as partial geometry. Against this background, we present a regular-sparse signature design based on the concept of Euler-squares. In Section IV, an iterative receiver design of low complexity is discussed, exploiting extrinsic information transfer between a MPA-based multi-user detector (MUD) and a bank of parallel FEC decoder. In Section V, we provide numerical examples illustrating the interplay between the relevant system parameters and compare the proposed scheme numerically against the state-of-the-art. Finally, Section VI concludes this paper.

Notation: If not stated otherwise, we use the following notation: scalars, vectors and matrices are denoted as $x$, $x$ and $X$, respectively. Throughout this paper, random variables (i.e. $X$) and particular realizations thereof (i.e. $x$) are denoted as non-italic letters. With a slight abuse of the notation, a multidimensional random variable is generally denoted as non-italic boldfaced letter $X$. For transpose- and Hermitian transpose the notation $\cdot^T$ and $\cdot^H$ is used. A set is denoted as $\mathcal{X}$, whereas the specific set of integers $\{1, 2, \ldots, N\}$ is denoted as $[N]$.

II. System Model and Coding Scheme

A. System Model

Consider a multiple access channel where a set $\mathcal{K}$ of $K$ users in total attempt to access a block of $N$ orthogonal resources (channel uses). If a user $k$ is active (scheduled or at random), a codeword $X_k \in \mathbb{C}^N$ is selected for transmission. We note that, in the case of scheduled access, the activity pattern is known to the receiver. In the grant-free scenario with random activation, on the other hand, we assume that the receiver knows only the statistics of $a$, but not its’ realization.

Considering a Gaussian channel model, the receiver observes

$$Y = \sum_{k \in \mathcal{K}} a_k \sqrt{P_k} X_k + W,$$ (2)
where \( \mathbf{a} = (a_1, a_2, \ldots, a_K) \in \{0, 1\}^K \) is a vector denoted as the "activity pattern"; \( P_k \) is the transmit power of user \( k \) (possibly also incorporating path-loss effects); the elements of \( \mathbf{W} \in \mathbb{C}^N \) are i.i.d. \( \mathcal{CN}(0, 1) \); and \( \mathbf{Y} \in \mathbb{C}^N \). The system model is illustrated in Fig. 3. In the following, we assume the same power level for all users, and define the SNR \( \Delta \triangleq P \) as \textit{per-user} signal-to-noise ratio.

**B. Block-sparse Coded Modulation**

The \( N \) channel uses are split into \( n_s \) sub-blocks of length \( \ell = N/n_s \). When active, user \( k \) transmits a complex vector (codeword) \( \mathbf{b}_k \in \mathbb{C}^\ell \), subject to the power constraint \( \|\mathbf{b}_k\|_2^2 \leq \ell \). The encoding process is as follows. When active, user \( k \) selects a message \( \omega_k \) from its message set \( \mathcal{W}_k \triangleq \{1, \ldots, M_k\} \) and maps it to a binary codeword \( \Delta_k \in \mathbb{F}_{2}^{n} \) from a FEC code \( \mathcal{C}_{FEC}^{(k)} \). While in general it is possible that each user applies a different FEC code (which can also be non-linear), in the following we will assume that all users apply the same linear block code \( \mathcal{C}_{FEC} \) of rate \( R_{FEC} = \log_2 |\mathcal{W}|/n \), where \( \mathcal{W} \triangleq \mathcal{W}_1 = \cdots = \mathcal{W}_K \) is the corresponding message set. Next, the encoded block of bits \( (\Delta_k) \) are interleaved \( \Delta'_k = \pi_k(\Delta_k) \) in order to randomize dependencies between bits. After the interleaver, in the modulation step, the binary codeword \( \Delta'_k \) is mapped to a complex vector \( \mathbf{b}_k \) at a modulation order \( Q_k = \ell/n \). The codeword \( \mathbf{b}_k \) is spread over the \( n_s \) sub-blocks by using a (sparse) signature \( \mathbf{s}_k \in \mathbb{C}^{n_s} \), resulting in the transmit codeword \( \mathbf{X}_k = \mathbf{s}_k \mathbf{b}_k^T \in \mathbb{C}^{n_s \times \ell} \) over the block of in total \( n_s \cdot \ell \) resources, as illustrated in Fig. 4.

We consider \textit{sparse} signatures \( \mathbf{s}_k \), chosen as columns of the matrix \( \mathbf{S} = \frac{1}{\sqrt{p}} \text{diag}(\phi) \mathbf{F} \), where \( \mathbf{F} \in \{0, 1\}^{n_s \times K} \) is a binary regular-sparse matrix, whose construction and properties
are discussed in detail in Section [III] and $\phi \in \mathbb{C}^K$ is a phase-rotation [25] on the complex union circle. Hence, our code achieves an energy-per-bit to noise power spectral density ratio 

$$E_b/N_0 \triangleq \frac{\ell}{2 \log_2 |\mathcal{W}|} \cdot \text{SNR per complex dimension} \quad \text{with effective code rate} \quad R = \frac{\log_2 |\mathcal{W}|}{n_s \cdot \ell} = \frac{R_{\text{FEC}}}{n_s} \text{ bits per channel use.}
$$

C. Application in an OFDM-based system

When applied over an orthogonal frequency division multiplex (OFDM)-based system, the $n_s$ sub-blocks in the above model may correspond to resource blocks (RBs), each consisting of $\ell = n_{sc} \cdot n_o$ time-frequency slots, where $n_o$ is the number of OFDM symbols spanning over $n_{sc}$ consecutive subcarriers. When active, user $k$ spreads its message over only a small number of sub-blocks $\rho \ll n_s$, where $\rho$ is the sparsity (i.e. the number of non-zero elements) of the signature $s_k$. In this case $n_o$ is related to the message duration, whereas $\rho \cdot n_{sc}$ is related to the bandwidth assigned to a given user. While we have so far considered a Gaussian channel model, we note that in general the RBs can model a block-fading scenario where the channel fading is assumed to stay constant within each RB, and to change independently from RB to RB. In that case the sparsity parameter $\rho$ represents the number of (frequency) diversity branches over which the users span their messages. We note that, in the block-fading scenario, a fraction of the $\ell$ channel resources within each RB should be dedicated for training, i.e. for pilot transmission (see [26] for a discussion on the effect of pilot design on the reliability performance in the finite blocklength regime).

$^2$Typical values, e.g. in Long-Term Evolution (LTE), are $n_{sc} = 12$ and $n_o = 14$. 

Fig. 4: Illustration of the coding scheme: the outer product of a finite-length codeword $b_k$ and a sparse signature $s_k$ constitute the block-sparse codeword $X_k$ of size $N = n_s \cdot \ell$. 
III. SPARSE CODE DESIGN FROM EULER SQUARES

In the following we present the details of the sparse code design based on the concept of Euler-Squares [21], and discuss the connection to code designs from partial geometries. The code construction is motivated by recent works on deterministic binary matrices for compressed sensing [22]. The construction in (1), which is derived from the Euler Square $E(3, 2)$ [21], has the following properties:

1) The mapping is bi-regular, such that the information from each user is mapped on exactly $\rho = 2$ resources, and exactly $\gamma = 3$ users overlap on each of the available resources.
2) Two rows (columns) of $F$ have at most one place where both have non-zero entries (i.e. there is an overlap on mostly one position);
3) $F^T$ is a $\gamma \times \rho$ array of $\gamma \times \gamma$ circulant permutation matrices (CPMs).

As we will discuss in more detail in Section III compared to general bi-regular mappings, property 2) and 3) may offer additional advantages for the decoding process both in the case of scheduled and grant-free transmissions. For example, from 2) it follows that any two users will overlap on one resource at most. Similarly, 3) provides connection to code constructions based on cyclic permutation matrices, providing compact protograph representations and offering insights to the decoding properties via the investigation of trapping sets (see [23] for a related discussion on the decoding properties of low-density parity check (LDPC) codes).

A. Graph-Theoretic Background

1) Bi-regular Graphs: A graph $G(V, E)$ with vertex set $V$ and edge set $E$ is denoted as bipartite, if its vertex set $V$ can be expressed as the union of two sets $U \cup W$ such that all edges of $E$ are between vertices in the sets $U$ and $W$. A walk of length $k$ in a graph $G$ is a sequence of vertices $(v_0, v_1, \ldots, v_k)$ such that any consecutive vertices $(v_i, v_{i+1})$, $\forall i \in \{0, 1, \ldots, k - 1\}$ form an edge in $E$. The walk is of length $k$ if it traverses $k$ edges. A walk is closed, if $v_k = v_0$. Further, a walk is a cycle, if the vertices of the walk are distinct (except for $v_k = v_0$) and cycle-free (CF), if it is a closed walk with no cycles. The girth $g$ of the graph is the length of the shortest cycle in $G$. A graph $G$ with vertex set $V$ can be represented as a $v \times v$ adjacency matrix $A$ with $[A]_{i,j} = 1$, if $(v_i, v_j) \in E$ and zero otherwise, where $|V| = v$. The set of (real) eigenvalues of $A$ is referred to as the spectrum of the graph. A $(k, n)$-bi-regular bipartite graph is a graph whose $u = |U|$ left (or variable) nodes have degree $k$ (incident with $k$ edges), and whose $w = |W|$ right (or check) nodes have degree $n$, while the number of edges is $uk = wn$. 
2) Partial Geometries: Consider a system composed of a set of points \( \mathcal{N} \) and a set of lines \( \mathcal{M} \) (in which a line is defined as a set or points). If a line \( l \in \mathcal{M} \) contains a point \( n \in \mathcal{N} \), we say that \( n \) is on \( l \) and \( l \) passes through \( n \). If two points are on a line, the two points are adjacent and if two lines pass through the same point these two lines intersect (otherwise they are parallel).

The system composed of the sets \( \mathcal{N} \) and \( \mathcal{M} \) is a partial geometry \( \text{PaG}(\gamma, \rho, \delta) \) \[23\], if the following conditions are satisfied for some fixed integers \( \gamma \geq 2, \rho \geq 2 \) and \( \delta \geq 1 \): i) any two points are on one line; ii) each point is on \( \gamma \) lines; iii) each line passes through \( \rho \) points; and iv) if a point \( n \) is not on line \( l \), there are exactly \( \delta \) lines, each passing through \( n \) and a point on \( l \).

B. Construction from Euler Squares

1) Existence of Euler Squares: An Euler square \( E(\gamma, \rho) \) of order \( \gamma \) and degree \( \rho \) is a square array of \( \gamma^2 \) \( \rho \)-tuple of numbers \((a_{ij1}, a_{ij2}, \ldots a_{ij\rho})\), where \( a_{ijr} \in \{0, 1, 2, \ldots, \gamma - 1\} \) with \( r = 1, 2, \ldots, \rho; i, j = 1, 2, \ldots, \gamma; \gamma > \rho; a_{ipr} \neq a_{iqr} \) and \( a_{pjr} \neq a_{qjr} \) for \( p \neq q \) and \( a_{ijr}a_{ijs} \neq a_{pqr}a_{pqs} \) for \( i \neq p \) and \( j \neq q \). Explicit constructions of Euler Squares are known to exist for the following cases \[21\]:

1) \( E(p, p - 1) \), where \( p \) is a prime number.
2) \( E(p^r, p^r - 1) \), where \( p \) is a prime number.
3) \( E(\gamma, \rho) \), where \( \gamma = 2^r p_1^{r_1} p_2^{r_2} \ldots p_l^{r_l} \) for distinct odd primes \( p_1, p_2, \ldots, p_l \), and \( \rho + 1 = \min \{2^r, p_1^{r_1}, p_2^{r_2}, \ldots, p_l^{r_l}\} \).

Furthermore, the existence of the Euler Square \( E(\gamma, \rho) \) implies that the Euler Square \( E(\gamma, \rho') \), with \( \rho' < \rho \), also exists. Based on the above, for \( \gamma \geq 3, \rho \geq 2 \), a binary matrix for sparse coded access \( F \) of size \( \gamma \rho \times \gamma^2 \) is constructed as:

\[
f_{ij} = \begin{cases} 
1 & \text{if } (a_{ij})_{\lfloor \frac{i-1}{\gamma} \rfloor + 1} \equiv (i-1) \mod \gamma \\
0 & \text{otherwise},
\end{cases}
\] (3)

where \((a_{ij})\) is the \( i \)-th \( \rho \)-tuple, \((a_{ij})_l\) is the \( l \)-th element in the \( j \)-th \( \rho \)-tuple and \( \lfloor x \rfloor \) denotes the largest integer not greater than \( x \). Consequently, \( F \) is effectively a block matrix consisting of \( \rho \) number of \( \gamma \times \gamma^2 \) blocks, where there are exactly \( \rho \) ones in each column of \( F \) and each column of \( F \) correspond to a \( \rho \)-ad in the Euler Square \( E(\gamma, \rho) \). We note that the example in \[1\] corresponds to a construction from the Euler Square \( E(3, 2) \).
C. Properties of the Construction

By construction (3), the matrix $F$ fulfills the following properties:
1) $F$ is bi-regular, with $\rho$ non-zero entries in each row, and $\gamma$ in each column;
2) Two rows (columns) of $F$ have at most one place where both have non-zero entries (i.e. there is an overlap on one position at most. Matrices of this type are called row and column (RC) constrained;
3) $F^T$ is a $\gamma \times \rho$ array of $\gamma \times \gamma$ CPMs.

1) Graphical Representation: Based on the above properties, following [23, Theorem 1], we observe that $F^T$ is the line-point incidence matrix of a partial geometry $PaG(\gamma, \rho, \rho - 1)$ with $n = \gamma \rho$ points corresponding to the columns of $F^T$ and $m = \gamma^2$ lines corresponding to the rows of $F^T$. The corresponding bipartite graph has $m = \gamma^2$ variable nodes (VNs) of degree $\rho$ (representing the $m$ lines in $PaG(\gamma, \rho, \rho - 1)$), and $n = \gamma \rho$ check nodes (CNs) of degree $\gamma$ (representing the $n$ points in $PaG(\gamma, \rho, \rho - 1)$). The associated partial geometry $PaG(\gamma, \rho, \rho - 1)$ is quasi-cyclic (QC), due to the quasi-cyclic structure of the associated line-point incidence matrix (see [23] for a general overview of quasi-cyclic partial geometries).

Remark: We note that, while the construction from the $E(\gamma, \rho)$ is a quasi-cyclic partial geometry QC - $PaG(\gamma, \rho, \rho - 1)$, not all quasi-cyclic partial geometries of the type QC - $PaG(s, t, t - 1)$ correspond to an Euler Square construction, since constructions of Euler Squares do not exist for all $s, t > 1$. In fact, Euler Squares constructions have additional structural properties, i.e. they represent a sub-class of the quasi-cyclic partial geometries QC - $PaG(\gamma, \rho, \rho - 1)$.

2) Protograph Representation: The matrix $F^T$ constitutes a binary $\gamma \times \rho$ array of CPMs of order $\gamma$ of the following form:

$$F^T = \begin{bmatrix}
C_{0,0} & C_{0,1} & \cdots & C_{0,\rho-1} \\
C_{1,0} & C_{1,1} & \cdots & C_{1,\rho-1} \\
\vdots & \vdots & \ddots & \vdots \\
C_{\gamma-1,0} & C_{\gamma-1,1} & \cdots & C_{\gamma-1,\rho-1}
\end{bmatrix},$$

where $C_{i,j}$ is uniquely specified by the location of the single 1-entry of its top row, called the generator. If the single 1-entry of the top row of $C_{i,j}$ is located at the position $k_{i,j}$, $0 \leq k_{i,j} < \gamma$, then we use $(k_{i,j})$ to specify the CPMs $C_{i,j}$.

Following [23], one can divide the VNs of $PaG(\gamma, \rho, \rho - 1)$ in $\gamma$ disjoint clusters $\Phi_0, \Phi_1, \ldots, \Phi_{\gamma-1}$, and the CNs in $\rho$ disjoint clusters, $\Omega_0, \Omega_1, \ldots, \Omega_{\rho-1}$ CNs. Consequently, one can construct a
(bipartite) protograph with $\gamma$ (super) VNs and $\rho$ (super) CNs. The protograph $\text{PaG}(\gamma, \rho, \rho - 1)$ contains all the structural information of the matrix $F$ derived from $E(\gamma, \rho)$. An example of a protograph associated with the construction from the Euler Square $E(3, 2)$ is depicted in Fig. 5.

$$\Phi_0 = \{v_0, v_1, v_2\}$$
$$\Phi_1 = \{v_3, v_4, v_5\}$$
$$\Phi_2 = \{v_6, v_7, v_8\}$$
$$\Omega_0 = \{c_0, c_1, c_2\}$$
$$\Omega_1 = \{c_3, c_4, c_5\}$$

Fig. 5: Protograph representation of the bipartite graph associated with the construction from $E(3, 2)$.

3) Implications on Encoding/Decoding: For a sparse signature design associated with a bipartite graph $G(V, C, E)$, its error performance depends on a number of structural properties of $G$. One of these properties is the girth, which is the length of the shortest cycle in $G$. For the bipartite graphs associated with the partial geometries $\text{PaG}(\gamma, \rho, \rho - 1)$ considered here, it has been shown in [23] that the girth is 8 when $\rho = 2$, and there are $\frac{1}{4} \gamma^2 (\gamma - 1)^2$ cycles of length 8 [27]. Further, for $\rho > 2$, the girth of the associated graph is 6 and there are $\frac{1}{6} \gamma^3 (\gamma - 1)(\rho - 2)(\rho - 1)$ cycles of length 6 [27]. Besides the impact on the decoding procedure, the girth size also provides a hint on the number of iterations that need to be performed in each MPA step of the turbo-decoding procedure. As also evident from the simulations, performing a number of message passing iterations beyond the girth does not bring additional benefits on average.

Another important property of $G$ is connectivity, defined as the number of VNs which are connected to a specific VN by paths of length 2, which has an effect on the rate of convergence of the message passing procedure. For the $(\gamma, \rho)$-bi-regular graphs of interest here, the connectivity is $\rho(\gamma - 1)$. In general, the message passing algorithm running on a bipartite graph with higher connectivity converges faster.

Besides the girth and the connectivity, an important factor is also the size of the so called trapping sets. The analysis of trapping sets for the graphs of interest here is beyond the scope of the paper. We note that the problem of determining the sizes of trapping sets for LDPC codes
obtained from finite geometries was initiated in [27]. In the special case of partial geometries, some bounds were presented in [23]. We note that the use of partial geometries to construct parity check matrices for binary LDPC code has resulted in the design of codes with excellent error performance. We note, that the quasi-cyclic structure present in the construction also allows for the implementation of efficient encoding algorithms. Furthermore, as discussed, due to the quasi-cyclic structure, the bipartite graph associated with QC – PaG(γ, ρ, ρ − 1) can be effectively represented by a much smaller protograph.

IV. RECEIVER PROCESSING

A. Turbo-based Iterative Receiver Design

The iterative decoding process is characterized by the exchange of soft information between the MUD module employing a MPA, which operates on the bipartite graph defined by the sparse code for accessing the channel, and the bank of individual FEC modules, as depicted in Fig. 6.

Fig. 6: Iterative receiver structure employing turbo-enhanced message passing algorithm (MPA) - based multi-user detector (MUD) and forward error correction (FEC) decoding.

The decoding process runs as follows (e.g. see [28]). In the first step, the MUD module delivers the a-posteriori log-likelihood ratio (LLR) of a transmitted "+1" and "-1" (considering binary phase shift keying (BPSK) modulation) for each user \( k \in [K] \), and each code bit \( i \in [\ell] \),
as

\[ \Lambda_k^1[i] \triangleq \log \frac{P(b_k[i] = +1 \mid y[i])}{P(b_k[i] = -1 \mid y[i])} \]
\[ \triangleq \log \frac{p(y[i] \mid b_k[i] = +1)}{p(y[i] \mid b_k[i] = -1)} + \log \frac{P(b_k[i] = +1)}{P(b_k[i] = -1)}, \]

(4)

where \( \lambda_k^{2p}[i] \) denotes the \textit{a-priori} information from the previous iteration step (which is zero in the first iteration) and \( \lambda_k^1[i] \) denotes the \textit{extrinsic} information provided by the MUD.

In the second step, the extrinsic information of the MUD is de-interleaved and fed into the \( k \)-th user’s channel decoder, as \textit{a-priori} information \( \lambda_k^{1p} \) in the next iteration step. The channel decoder then calculates \textit{a-posteriori} LLR of each code bit for each user independently, as

\[ \Lambda_k^2[i] \triangleq \log \frac{P(b_k[i] = +1 \mid \lambda_k^{1p}; \text{decoding})}{P(b_k[i] = -1 \mid \lambda_k^{1p}; \text{decoding})} \]
\[ \triangleq \lambda_k^2[i] + \lambda_k^{1p}[i], \]

(5)

which is again the sum of the \textit{prior} (\( \lambda_k^{1p} \))- and the \textit{extrinsic} information (\( \lambda_k^2[i] \)).

As discussed in the Introduction, the expected advantage of using sparse signature for channel access in combination with message passing (i.e. BP)-based detection is that, under certain conditions, BP yields an equivalent scalar Gaussian channel from the viewpoint of each user, on which per-user FEC operates. In the finite-size regime of interest, the two aspects, namely sparse code design for channel access and FEC, should be addressed jointly.

\textbf{B. Random User Activation}

With random activation of users, both the multiuser performance and the receiver complexity is determined by the pruned bipartite graph which contains only the variable nodes associated with the active users. In this context, it is important to determine under which choices of the full graph (i.e. the signature matrix), the pruned graph can lead to successful decoding with high probability for arbitrary random activation patterns. An example is depicted in Fig. 7, where the empirical factor-node degree distribution of the pruned - and the residual factor graph (after peeling decoding) is depicted for a construction based on \( E(101, 2) \) for different \( K_a \).

Depending on the connectivity of the sparse bipartite graph, the received signals (observations) associated with the check nodes can be categorized into the following types: (i) \textit{zero-ton} check
Fig. 7: Empirical factor node degree distribution before (dashed) and after (solid) peeling-decoding (removing singletons) for the bipartite graph based on $E(101, 2)$ at different numbers of active devices $K_a$, being active at random.

node which does not involve any non-zero symbols; (ii) single-ton check node which involves only one non-zero symbol; (iii) multi-ton check node whose value is the sum of more than one non-zero symbol. In the following we provide more details on the peeling-based decoding procedure.

Consider a "genie" that informs the decoder about the set of active users $K_a$, i.e. about the structure of the pruned graph (which parity check nodes are zero-ton, single-ton and multi-ton). In the following, let $\mathcal{V}[t]$ denote the set of variable nodes remaining at peeling step $t$, and let $\mathcal{C}[t]$ denote the set of check nodes remaining at peeling step $t$. Further, let $\mathbf{F}[t] = \mathbf{F}(\mathcal{V}[t])$ denote the corresponding incidence matrix. The peeling process starts with the initialization $\mathcal{V}[0] = K_a$. Now, let $\mathcal{C}^s[t] \subseteq \mathcal{C}(t)$ denote the set of single-ton check nodes at step $t$, and let $\mathcal{V}^s[t] \subseteq \mathcal{V}(t)$ denote the subset of variable nodes (i.e. users) that are connected to at least one single-ton check node. If $\mathcal{U}^s[t] \neq \emptyset$, repeat:

(I) Decode the messages $\hat{b}_k, \forall k \in \mathcal{U}^s[t]$.

(II) Peel-off the estimated messages from the received signal

$$Y^r[t + 1] = Y^r[t] - \sum_{k \in \mathcal{U}^s[t]} s_k \hat{b}_k^T.$$  \hspace{1cm} (6)

(III) Update

a) $\mathcal{U}[t + 1] = \mathcal{U}[t] \setminus \mathcal{U}^s[t]$

b) $\mathbf{F}[t + 1] = \mathbf{F}(\mathcal{U}[t + 1])$

c) $\mathcal{C}^s[t + 1]$ from $\mathbf{F}[t + 1]$.

(IV) If $\mathcal{C}^s[t + 1] \neq \emptyset$: set $t = t + 1$ and proceed with (I).
When there is no genie, the pruned graph (i.e., equivalently, the set of active users $K_a$) needs to be estimated in some way and this is the realistic case under which the algorithm operates. This problem is essentially a one of sparse support recovery in compressed sensing for which theoretical analysis has been presented in [29], where a scheme based on sparse graph codes was proposed. The scheme provably achieves an order-optimal scaling in both the measurement cost and the computational run-time in the sub-linear sparsity regime. Hence, using the peeling-decoder for the massive multiple access requires a second stage to decode all messages associated with the residual factor graph, especially if the number of active devices increases.

In our context, these insights may be used to modify the sparse resource element mapping such as to ensure reliable recovery of the structure of the pruned graph. Following [29], in order to obtain the information required for the peeling decoding process, one can assign a vector-weighted complex sum to the check nodes [29], where each variable node $x_k$ (associated to the $k$-th UE) is weighted by the $k$-th column of a matrix $\tilde{S}$. In the noiseless case, for example, $\tilde{S}$ can be selected as the first two rows of a discrete Fourier transform (DFT)-matrix

$$
\tilde{S} = \begin{bmatrix}
1 & 1 & 1 & 1 & \ldots & 1 \\
1 & W & W^2 & W^3 & \ldots & W^{K-1}
\end{bmatrix},
$$

(7)

where $W = e^{j \frac{2\pi}{K}}$. Translated to our scenario with random user activation, this means that the support recovery requires reservation of $n_s$ resources out of the $N = \ell \cdot n_s$ resources in total, resulting in an effective decrease of the spectral efficiency $\eta$ by a factor of $N - O(\frac{K_a \log(K)}{N})$.

While the analysis in the noisy setting is somewhat more involved and, in general depends on the SNR, in [29] it has been shown that, when the sparse coefficients take values in a finite and quantized alphabet, the framework can achieve a robust estimation of the support using $O(K_a \log(K/K_a))$ measurements obtained from measurement matrix with elements $\{-1, 0, 1\}$, where we recall that $K$ is the total number of system users. By considering this value to be the overhead (in terms of channel resources) associated with the recovery of the active set, one can (roughly) claim an effective decrease of the spectral efficiency $\eta$ by a factor of $\frac{N - n_s}{N}$.

**Discussion:** We note that this overhead will be moderate in the scenario of interest where the expected number of active users is sub-linear in the total number of system users, $K_a = O(K^\delta)$, for some $0 < \delta < 1$, but the number of available resources scales linearly with the total number of system users, $N = O(K)$ with fixed system overload $K/N$. An optimal design of the measurement process for the estimation of the active set of users is out of the scope of this
paper and is left for future consideration. For the numerical simulations we will assume that a sufficiently large fraction of the available resources is reserved for the estimation of the active set (i.e. the pruned graph structure), such that reliability of this estimation step meets a required level.

V. PERFORMANCE EVALUATION

In this section, we provide numerical results to illustrate the key properties of the transmission scheme and the receiver processing architecture. The proposed coding scheme, and specifically the sparse signature construction based on $E(\gamma, \rho)$ (as discussed in Section III) is flexible in the sense that it can be explicitly described/configured for a wide number of system parameters, i.e. number of users $K = \gamma^2$; signature length $n_s = \gamma \cdot \rho$; sparsity (number of non-zero elements) of the signature $\rho$; and block-size $\ell = N/n_s$, where $N$ denotes the overall number of channel uses. First, we compare the spectral efficiency of the proposed regular-sparse constructions to the analytic (quasi-) closed-form expression for regular constructions in the LSL.

A. Spectral Efficiency

It was shown in [16] for the Gaussian channel model that in the LSL, which assumes infinitely long signatures $n_s \to \infty$, regular-sparse signature constructions outperform both irregular-sparse and dense constructions in terms of their spectral efficiency (defined as the total achievable throughput in bits per channel use). In Fig. 8 we quantify the spectral efficiency of different Euler-squares based constructions relative to the LSL semi-analytical expressions from [16].

Based on the system model in (2), for the constructions from Euler squares, the spectral efficiency is evaluated as

$$C_{\text{Opt}}^{n_s}(S) = \frac{1}{n_s} \sum_{i=1}^{n_s} \log_2 \left( 1 + \text{SNR} \cdot \lambda_i \left( SS^T \right) \right),$$  (8)

where $\lambda_i(X)$ corresponds to the $i^{th}$ eigenvalue of $X$.

We observe that the constructions from Euler-squares, although having finite signature size, perform fairly well over a range of values of the ratio $\beta = K/n_s$. This, together with the low-complexity MUD based on message passing, motivates the use of sparse deterministic signature for non-orthogonal access, both in the scheduled and in the grant-free scenario.
Fig. 8: Spectral efficiency per dimension for different regular constructions $E(\gamma, \rho)$ at $E_b/N_0 = 10$ dB compared to the asymptotic results [16], denoted as $C^{LSS}$ where $d$ denotes the number of non-zero elements.

**B. Convergence of the Iterative Decoding**

The receiver operates in an iterative fashion exchanging information between a joint MPA - stage (to decouple the individual user-signals) and a user-specific FEC stage, to combat noise and provide extrinsic information to exploit turbo-decoding principle, as depicted in Fig. 9. The (MPA) - stage performs BP on the factor - graph, provided by the underlying Euler-square construction. The (FEC) - stage is implemented using a bank of legacy LDPC decoder generates user-specific soft information which is fed back as extrinsic information every fixed number of iterations.

![Diagram](image-url)
As a tool for the analysis, we use the scattered extrinsic information transfer (S-EXIT) charts, originally proposed in [30] as a guideline to design finite length LDPC codes. In the context of our iterative receiver architecture, the S-EXIT chart uses the statistics of numerous decoding trajectories obtained from simulations of the actual joint MUD and the channel decoder, and tracks their frequency of occurrence in a two-dimensional histogram over the EXIT mutual information plane. Thus, it enables to get an insightful expectation of the iterative decoding behavior of the code enabling statistical analysis on several instances of it. The mutual information between an observation LLR value $\Lambda$ and transmitted code-words $b$ is approximated as [30]:

$$I(\Lambda; b) = 1 - \mathbb{E} \left[ \log_2 \left( 1 + e^{-\Lambda} \right) \right]$$

$$\approx 1 - \frac{1}{K} \sum_{k=1}^{K} \left( \frac{1}{\ell} \sum_{i=1}^{\ell} \log_2 \left( 1 + e^{-\Lambda_k[i] - b_k[i]} \right) \right).$$

In Fig. 10, the extrinsic information exchange between the MPA - and the FEC module is depicted for different parameter of the proposed code constructions. Each grey line represent the vertice of a decoding- process for a given Euler-square construction where the random variable for each realization is the noise and the FEC generator. It is observed, that the output of the MPA / FEC, i.e. the edges of the decoding trajectory (illustrated as red / blue points in Fig. 10), are clustered in groups, where each cluster represents one feedback iteration in the joint decoding process, illustrated in Fig. 9. In general, we observe from Fig. 10 that the decoding process is characterized by (i) a low number of iterations of feedback between the modules to reach the point $(1, 1)$, which is in the order of $\approx 3$ for $E(7, 2)$ and $\approx 2$ for $E(7, 3)$ and (ii) a sufficient separation of the MPA and the FEC decoding cluster for the decoding trajectory to "sneak" through the bottleneck [30].

Fig. 10: S-EXIT-chart, illustrating the extrinsic information exchange between the MPA and FEC - module for different parameter of code constructions.
C. Empirical Results

In the following, we provide empirical performance results of the proposed coding scheme when applied in different massive access access protocols [31], namely (i) grant-based non-orthogonal multiple access; (ii) grant-free (massive) random access, where only a small subset of users are active at the same time and (iii), we discuss our scheme in the context of unsourced random access. The figure of merit is the average error probability, defined as

\[ P_e = \frac{1}{K} \sum_{k \in K} \Pr \{ w_k \neq \hat{w}_k \mid w_k \}, \]

(11)

where \(|K| = K\) are the number of active devices. Further, we refer to the total number of bits per channel use which need to be decoded at the receiver, as spectral efficiency, defined as

\[ \eta = \frac{K \cdot k}{N}. \]

(12)

The encoding is done as described in Section II-B. Throughout the following analysis, if not stated otherwise, we consider a regular LDPC code as FEC with parity-check matrix generated following Gallager’s algorithm [32].

1) Grant-based Non-Orthogonal Multiple Access: First, we consider a non-orthogonal multiple access scenario, where users are co-scheduled on the same physical resources. The activity pattern as well as the seed of the FEC and the FEC per device can be assumed to be known at the receiver. In this setting, the proposed construction \( E(\gamma, \rho) \) provides the flexibility to trade different system parameter, e.g. overload and message size. In this context, we are specifically interested in characterizing the system in the finite-block length regime, i.e. \( n_s \cdot \ell \ll \infty \), with the objective to increase the spectral efficiency, i.e. the number of users sharing each resource element \( \beta = \gamma/\rho \). The trade-off between the channel coding gain (as a function of the FEC blocklength \( \ell \)) and the system overload (as a property of the specific Euler-square construction) \( \beta \) is illustrated Fig. [11] where the error rate performance for different Euler-square constructions is plotted, namely \( E(5, 2) \) and \( E(5, 4) \) in conjunction with different message sizes \( k = \{32, 62, 122\} \) at \( R_{FEC} \approx 0.5 \). Note, that both constructions \( E(5, 2) \) and \( E(5, 2) \) accommodate \( K = 25 \) devices at an overload of 2.5 and 1.25, respectively. Hence, all settings in Fig. [11] are operated at the same target spectral efficiency. constructions provide the same spectral efficiency.

From Fig. [11] one can observe the following: (i) a larger message size improves the performance in terms of \( P_e \) in the lower range of the water-fall region, for both constructions. This effect is directly connected to the higher coding gain which comes with the larger FEC
Fig. 11: Error vs $E_b/N_0$ for different code constructions and message sizes ($k$) at FEC rate $R_{FEC} \approx 0.5$.

blocklength; (ii) for a fixed number of resources ($N$), the system operated at higher overload ($E(5, 2)$) requires a higher energy-per-bit to achieve a similar performance compared to $E(5, 4)$.

To further analyze this effect, Fig. 12 plots the required energy-per-bit $E_b/N_0$ to achieve a target error of $P_e \leq 0.05$ as a function of the overload, which is achieved by using different Euler-square constructions in combination with an FEC of rate $R_{FEC} \approx 0.5$ at different message sizes.

We can observe two effects from Fig. 12: (i) the cost in terms of energy-per-bit only increases moderately with the overload in the observed range and (ii) that a larger message size improves the energy efficiency due to the higher coding gain of the FEC.

For example, at message size $k = 32$, the required energy-per-bit increases from 7.3 dB at $\beta = 1.25$ to 7.6 dB at $\beta = 2.5$ to achieve a similar performance in terms of $P_e$. Further, the higher coding efficiency, as an effect of a larger FEC blocklength, increases the energy efficiency by $\approx 0.6$ dB when the blocklength is increase by a factor of 2.

Discussion: For the grant-based non-orthogonal access protocol, the proposed scheme allows to trade-off system parameter like overload and energy-efficiency. For example, by choosing different Euler-square constructions (tailored to the targeted information message length), one can increase the spectral efficiency by a factor of 2 at the cost of $\approx 0.3$ dB per device for the investigated regime.
2) Grant-free Random Access: Consider a (massive) grant-free random access protocol, where only a small subset $K_a \subset K$ with $K_a/K \ll 1$ are active simultaneously and the activity pattern is unknown to the receiver. In this regime typically the total number of devices $K$ connected to the network is in the order of the number of available resources $N$ (or higher). Let $K_a^t \subset K$ be a random subset of active devices at slot $t$. With $|K_a^t| = K_a$, we refine the error probability (11) as

$$P_e = \lim_{T \to \infty} \frac{1}{T} \cdot \sum_{t \in [T]} \sum_{k \in K_a} \Pr \{ w_k \neq \hat{w}_k \mid w_k \}. \quad (13)$$

In the following, we assume that a sufficiently large fraction of the available resources is reserved for the estimation of singletons, such that peeling decoding can be applied, as discussed in Section IV-B. In Fig. 13, the error rate $P_e$ is plotted as a function of $E_b/N_0$ for different number of active devices $K_a$. In this setting, the coding scheme is based on $E(73, 2)$ in conjunction with FEC at $R_{FEC} \approx 0.5$ and payload size $k = 101$. Based on this configuration, in total $K = 5,329$ devices are supported at a total number of $N = 28,908$ resources.

It can be observed, that the cost in terms of energy-per-bit for increasing the number of active devices from $K_a = 50$ to $K_a = 250$ to achieve an error $P_e \leq 0.05$ is $\approx 0.6$ dB.

3) Evaluation in the Context of Unsourced Random Access: In the following, we discuss the performance of the proposed scheme in the context of the U-RA model \cite{9}. In this context, our
scheme can be modified to fit the U-RA setting by letting the active users choose randomly a sparse sequence from a shared codebook (i.e. a set of sparse signatures obtained from an Euler square construction) and an associated inteleaver pattern, in combination with a finite (short) blocklength FEC code. Different to other U-RA approaches, in our setting the joint MPA and FEC decoding at the receiver decouples the received signal into parallel channels, yielding a decoding performance that is essentially determined by the structure of the (finite-length) FEC code.

In Fig. 14, we evaluate the performance of our scheme in the U-RA setting (in the context of the mentioned recent works). As in [19] (denoted as "C&F+BAC") and [20] (denoted as "SPARC+AMP"), we plot the $E_b/N_0$ (as a function of the the number of active devices $K_a$) required to achieve a target error probability $P_e \leq 0.05$, with a (per device) message size of $k \approx 100$ bits, and a total number of $N \approx 30,000$ resources shared by all $K$ devices. In this setting, we consider different code configurations with sparse signatures constructed from the Euler-square designs $E(73, 2)$, $E(97, 2)$ and $E(113, 2)$ respectively and adjusted FEC code rate accordingly. The results show that the required energy per bit (per device) increases by $\approx 0.7$ dB for the configuration with $E(73, 2)$ and $\ell = 198$, when the number of active devices increases from 50 to 300. The configuration using signatures $E(97, 2)$ and $\ell = 153$, on the other hand,
Fig. 14: Required $E_b/N_0$ at target error probability $P_e \leq 0.05$ as a function of the number of active devices $K_a$, for different code configurations at payload $k = 100$ bit.

requires only $\approx 0.2$ dB increase in the required energy.

Importantly, the results indicate that the signature size can be traded with the FEC coding rate to optimize the performance depending on the system load (the number of active users in the system), especially when the number of active devices exceeds $K_a \geq 300$, where recent U-RA schemes show a rapid increase in required energy [33]–[36].

D. Summary of the Results and Discussion

While being far from extensive, the presented simulation results illustrate the potential of the proposed approach as a viable solution for massive non-orthogonal access, both for scheduled (i.e. grant based) and grant-free transmissions. Here, key roles play (i) the regular sparse signature design (i.e. the structure of the induced bipartite graph on which the message passing procedure for user separation takes place), (ii) the finite blocklength FEC code applied on the level of individual users, and (iii) the iterative decoding procedure where the MPA - and the FEC module exchange soft information to perform joint decoding of the overlapping users.

As illustrated by the S-EXIT charts in Fig. 9, given an appropriate parameterization of the sparse signature design and the FEC code, the extrinsic information transfer taking place between the MPA and the FEC decoding modules yields a sufficient separation of the MPA and the FEC
decoding cluster for the the decoding trajectory to "sneak" through the bottleneck and reach the point \((1,1)\), which shows to be robust in the finite blocklength regime under different configurations. We note that, in this context, the S-EXIT charts framework can be further employed to optimize the joint parameterization of the sparse signature design and the FEC code, which we leave for future research.

When applied in the grant-free random access setting, we illustrated empirically that the proposed construction allows for peeling decoding. In that case both the multiuser performance and the receiver complexity is determined by the pruned bipartite graph that contains only the variable nodes associated with the active users. In this context, the results illustrate that for the considered code parameterizations, and in the considered regime (number of active users), the pruned graph leads to successful decoding with high probability for arbitrary random activation patterns.

Following other works on massive access, we have mostly evaluated the performance of the proposed scheme in terms of the required (per-user) \(E_b/N_0\) to achieve a nominal target error probability (typically in the order of \(10^{-2}\) when targeting massive connectivity in the context of IoT applications.) In this setting, we observed that the \(E_b/N_0\) performance of the proposed construction is quite robust to the increase of the overload factor \(\beta\) in the case of scheduled transmissions, i.e. on the number of active users \(K_a\) in the grant-free random access setting, as illustrated by Fig. 12 and Fig. 13 respectively.

Finally, we also elaborated on the performance of the proposed scheme in the unsourced random access setting. In that context, we have modified our scheme by letting the active users choose randomly a sparse sequence from a shared set of sparse signatures, together with an associated inteleaver pattern. The performed numerical simulations in Fig. 14 illustrate the potential (in terms of energy-efficiency) of the proposed scheme in the scenario with higher user activation (e.g. beyond \(K_a = 300\) active users in the setting where the system users send fixed messages of \(k = 100\) bits over \(N = 30,000\) channel resources), compared to other approaches that show a rapid increase in the required \(E_b/N_0\) in this regime. We note, that in contrast to [9], the number of active devices \(K_a\) does not need to be known perfectly in our scheme, hence enabling true random access for U-RA.

On a final note, we remark that the performed numerical simulations also shed light on the discussion regarding the context in which the many-access and the unsourced random access channel model should be applied. Along these lines, our approach is applicable to both settings,
depending on whether the sparse signatures are user-specific and are associated with the users upon system registration, or the users randomly select a sparse signature from a shared signature pool upon transmission. In the first setting, the sparse signatures carry information about the identity of the users, and the number of system users $K$ is a parameter whose choice affects the error probability. In the second setting, on the other hand, the total number of users in the system does not affect the error probability, but if user activity detection is needed, part of the payload should contain identifying information for the users.

VI. CONCLUSIONS AND FUTURE WORK

In this paper we investigated the non-asymptotic performance of deterministic sparse signature designs for non-orthogonal and massive access in conjunction with finite-block error-correction coding. The sparse signature design was based on the concept of Euler squares, and was motivated by recent works on deterministic binary matrices for compressed sensing and on error-correcting codes based on partial geometries. The sparse regular construction gives signatures with small density, which support decoding algorithms with low computational complexity. The proposed construction can be explicitly characterized for a large number of combinations of system parameters, thus providing a framework for accommodating both (non-orthogonal) scheduled and massive (grant-free) random access. We described a transceiver chain with extrinsic information exchange between MPA-based MUD and user specific FEC based on LDPC, and evaluated the performance numerically in various simulation settings. In the grant-free scenario with random activation, we demonstrated the potential of peeling-based decoding to further decrease the receiver-complexity, which allowed to operate in the regime where the number of system users in the order of tens of thousands, with several hundreds being simultaneously active. While here we have resorted on LDPC codes for FEC, for future work it would be interesting to consider other code constructions for the finite blocklength regime, such as, for example, polar codes and algebraic codes. Also, it would be of interest to investigate the performance of the scheme in the presence of receive diversity, both in a multi-antenna receiver scenario and in a multi-cell Cloud/Fog Radio Access Network architecture. Finally, we note that the approach is also amenable for grant-free access in the context of URLLC, where further investigation is needed in the appropriate code parameterization and the joint optimization of the sparse signature design and the FEC code.
REFERENCES

[1] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, “A Survey on Non-Orthogonal Multiple Access for 5G Networks: Research Challenges and Future Trends,” *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2181–2195, Oct. 2017.

[2] S. Shamai, “A Broadcast Strategy for the Gaussian slowly fading Channel,” in *1997 IEEE International Symposium on Information Theory (ISIT)*, Ulm, Germany, Jun. 1997, pp. 150–.

[3] R. Ahlswede, “Multi-way Communication Channels,” in *Second International Symposium on Information Theory (Sep. 1971)*, Tsahkadsor, Armenian S.S.R., 1973, pp. 23–52.

[4] N. Abramson, “The ALOHA System: Another Alternative for Computer Communications,” in *Fall Joint Computer Conference (AFIPS â ˘A ´Z70)*. Houston, Texas: ACM, Nov. 1970, p. 281â ˘A¸ S285.

[5] P. Mathys, “A Class of Codes for a T Active Users out of N Multiple-Access Communication System,” *IEEE Trans. Inf. Theory*, vol. 36, no. 6, pp. 1206–1219, Nov. 1990.

[6] I. Bar-David, E. Plotnik, and R. Rom, “Forward Collision Resolution - a Technique for Random Multiple-Access to the Adder Channel,” *IEEE Trans. Inf. Theory*, vol. 39, no. 5, pp. 1671–1675, Sep. 1993.

[7] E. Paolini, C. Stefanovic, G. Liva, and P. Popovski, “Coded Random Access: Applying Codes on Graphs to Design Random Access Protocols,” *IEEE Commun. Mag.*, vol. 53, no. 6, pp. 144–150, Jun. 2015.

[8] X. Chen and D. Guo, “Many-Access Channels: The Gaussian Case with Random user Activities,” in *2014 IEEE International Symposium on Information Theory (ISIT)*, Honolulu, Hawaii, Jun. 2014, pp. 3127–3131.

[9] Y. Polyanskiy, “A Perspective on Massive Random-Access,” in *2017 IEEE International Symposium on Information Theory (ISIT)*, Aachen, Germany, Jun. 2017, pp. 2523–2527.

[10] R. Hoshyar, F. P. Wathan, and R. Tafazolli, “Novel Low-Density Signature for Synchronous CDMA Systems over AWGN Channel,” *IEEE Trans. Signal Process.*, vol. 56, no. 4, pp. 1616–1626, Apr. 2008.

[11] H. Nikopour and H. Baligh, “Sparse Code Multiple Access,” in *2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, London, United Kingdom, Sep. 2013, pp. 332–336.

[12] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, “Non-Orthogonal Multiple Access (NOMA) for Cellular Future Radio Access,” in *2013 IEEE 77th Vehicular Technology Conference (VTC Spring)*, Dresden, Germany, Jun. 2013, pp. 1–5.

[13] G. Liva, “Graph-Based Analysis and Optimization of Contention Resolution Diversity Slotted ALOHA,” *IEEE Trans. Commun.*, vol. 59, no. 2, pp. 477–487, Feb. 2011.

[14] M. Yoshida and T. Tanaka, “Analysis of Sparsely-Spread CDMA via Statistical Mechanics,” in *2006 IEEE International Symposium on Information Theory (ISIT)*, Seattle, Washington, Dec. 2006, pp. 2378–2382.

[15] S. Verdú and S. Shamai, “Spectral Efficiency of CDMA with Random Spreading,” *IEEE Trans. Inf. Theory*, vol. 45, no. 2, pp. 622–640, Mar. 1999.

[16] O. Shental, B. M. Zaidel, and S. Shamai, “Low-Density Code-Domain NOMA: Better be Regular,” in *2017 IEEE International Symposium on Information Theory (ISIT)*, Aachen, Germany, Jun. 2017, pp. 2628–2632.

[17] M. T. P. Le, G. C. Ferrante, T. Q. S. Quek, and M. Di Benedetto, “Fundamental Limits of Low-Density Spreading NOMA With Fading,” *IEEE Trans. Wireless Commun.*, vol. 17, no. 7, pp. 4648–4659, Jul. 2018.

[18] D. Guo and C.-C. Wang, “Multiuser Detection of Sparsely Spread CDMA,” *IEEE J. Sel. Areas Commun.*, vol. 26, pp. 421–431, 05 2008.

[19] O. Ordentlich and Y. Polyanskiy, “Low Complexity Schemes for the Random Access Gaussian Channel,” in *2017 IEEE International Symposium on Information Theory (ISIT)*, Aachen, Germany, Jun. 2017, pp. 2528–2532.
[20] A. Fengler, P. Jung, and G. Caire, “SPARCs and AMP for Unsourced Random Access,” in 2019 IEEE International Symposium on Information Theory (ISIT), 2019, pp. 2843–2847.

[21] H. F. MacNeish, “Euler Squares,” Annals of Mathematics, vol. 23, no. 3, pp. 221–227, Mar. 1922, Princeton University.

[22] R. R. Naidu, P. Jampana, and C. S. Sastry, “Deterministic Compressed Sensing Matrices: Construction via Euler Squares and Applications,” IEEE Trans. Signal Process., vol. 64, no. 14, pp. 3566–3575, Jul. 2016.

[23] Q. Diao, J. Li, S. Lin, and I. F. Blake, “New Classes of Partial Geometries and their Associated LDPC Codes,” IEEE Trans. Inf. Theory, vol. 62, no. 6, pp. 2947–2965, Jun. 2016.

[24] C. Boyd, Vehkalahti, and O. Roope; Tirkkonen, “Grant-Free Access in URLLC with Combinatorial Codes and Interference Cancellation,” in 2018 IEEE Globecom Workshops (GC Wkshps), Abu Dhabi, United Arab Emirates, Dec. 2018.

[25] M. Taherzadeh, H. Nikopour, A. Bayesteh, and H. Baligh, “SCMA Codebook Design,” in 2014 IEEE 80th Vehicular Technology Conference (VTC2014-Fall), Vancouver, BC, Canada, Dec. 2014, pp. 1–5.

[26] G. Durisi, T. Koch, and P. Popovski, “Toward Massive, Ultrareliable, and Low-Latency Wireless Communication With Short Packets,” Proceedings of the IEEE, vol. 104, no. 9, pp. 1711–1726, Sep. 2016.

[27] S. J. Johnson and S. R. Weller, “Codes for Iterative Decoding from Partial Geometries,” IEEE Trans. Commun., vol. 52, no. 2, pp. 236–243, Feb. 2004.

[28] X. Wang and H. V. Poor, “Iterative (Turbo) Soft Interference Cancellation and Decoding for Coded CDMA,” IEEE Trans. Commun., vol. 47, no. 7, pp. 1046–1061, Jul. 1999.

[29] X. Li, D. Yin, S. Pawar, R. Pedarsani, and K. Ramchandran, “Sub-Linear Time Support Recovery for Compressed Sensing Using Sparse-Graph Codes,” IEEE Trans. Inf. Theory, vol. 65, no. 10, pp. 6580–6619, Oct. 2019.

[30] M. Ebada, A. Elkelesh, S. Cammerer, and S. ten Brink, “Scattered EXIT charts for finite length LDPC code design,” in 2018 IEEE International Conference on Communications (ICC). IEEE, 2018, pp. 1–7.

[31] X. Chen, D. W. K. Ng, W. Yu, E. G. Larsson, N. Al-Dhahir, and R. Schober, “Massive Access for 5G and Beyond,” arXiv e-prints, p. arXiv:2002.03491, Feb. 2020.

[32] R. Gallager, “Low-Density Parity-Check Codes,” IRE Transactions on Information Theory, vol. 8, no. 1, pp. 21–28, Jan. 1962.

[33] A. K. Pradhan, V. K. Amalladinne, K. R. Narayanan, and J.-F. Chamberland, “Polar Coding and Random Spreading for Unsourced Multiple Access,” arXiv e-prints, p. arXiv:1911.01009, Nov. 2019.

[34] E. Marshakov, G. Balitskiy, K. Andreev, and A. Frolov, “A Polar Code Based Unsourced Random Access for the Gaussian MAC,” in IEEE Vehicular Technology Conference (VTC2019-Fall), Honolulu, Hawaii, USA, Sep. 2019, pp. 1–5.

[35] V. K. Amalladinne, A. K. Pradhan, C. Rush, J.-F. Chamberland, and K. R. Narayanan, “On Approximate Message Passing for Unsourced Access with Coded Compressed Sensing,” arXiv e-prints, p. arXiv:2001.03705, Jan. 2020.

[36] A. Fengler, P. Jung, and G. Caire, “SPARCs for Unsourced Random Access,” arXiv e-prints, p. arXiv:1901.06234, Jan. 2019.