

CHIRRRUP: a practical algorithm for unsourced multiple access

Robert Calderbank and Andrew Thompson

September 25, 2019

Abstract

Unsourced multiple access abstracts grantless simultaneous communication of a large number of devices (messages) each of which transmits (is transmitted) infrequently. It provides a model for machine-to-machine communication in the Internet of Things (IoT), including the special case of radio-frequency identification (RFID), as well as neighbor discovery in ad hoc wireless networks. This paper presents a fast algorithm for unsourced multiple access that scales to $\mathcal{C} = 2^{100}$ (active or non-active) devices (arbitrary 100 bit messages). The primary building block is multiuser detection of binary chirps which are simply codewords in the second order Reed Muller code. The chirp detection algorithm originally presented by Howard et al. [1] is enhanced and integrated into a peeling decoder designed for a patching and slotting framework. In terms of both energy per bit and number of active devices (number of transmitted messages), the proposed algorithm is within a factor of 2 of state of the art approaches. A significant advantage of our algorithm is its computational efficiency. We prove that the worst-case complexity of the basic chirp reconstruction algorithm is $\mathcal{O}[nK(\log_2^2 n + K)]$, where n is the codeword length and K is the number of active users. Crucially the complexity is sublinear in \mathcal{C} , which makes the reconstruction computationally feasible — a claim we support by reporting computing times for our algorithm. Our performance and computing time results represent a benchmark against which other practical algorithms can be measured. Unsourced multiple access, chirp codes, compressed sensing, reconstruction algorithms.

Glossary of notation

B	Number of bits per message.
\mathcal{C}	Total number of devices/messages, equal to 2^B .
K	Number of active devices or transmitted messages.
s	Activity pattern to recover.
x_i	Codeword corresponding to message i .
X	Codebook matrix whose columns are $\{x_i\}$.
z	Additive noise.
y	Measurement vector.
$\{\phi_{P,b}\}$	Binary chirp with matrix P and translate b .
P	$m \times m$ binary symmetric matrix.
b	Translate in \mathbb{Z}_2^m .
Q	Transmitted power per message.
$\mathcal{H}(\cdot)$	Walsh-Hadamard transform.
c	Number of children in the tree search enhancement of CHIRRAP.
2^p	Number of slots.
2^r	Number of patches.
\tilde{n}	Codeword length in each slot, equal to 2^m for some m .
n	Overall codeword length, equal to 2^{m+p+r} .
n	Codeword length, equal to 2^m for some m .
l	Vector giving number of parity check bits appended to each patch.
E_i	Error event associated with message i .
\mathcal{P}	Per-user probability of error.
ϵ	Error threshold.
Λ	Set of transmitted messages.
Γ	Set of recovered messages.

1 Introduction

One of the opportunities that justifies the development of a new generation of wireless networks (5G) is machine-to-machine (M2M) communication (see [2]). This Internet of Things (IoT) will support grantless, simultaneous communication of a large network of devices, each of which can be thought to transmit infrequently. In this paradigm the messages are small and the number of users is massive, and it is the messages rather than the identities of the users that must be recovered. The development of next generation wireless networks is guided by information theory which provides performance bounds that govern transmission of short packets (see [3]).

1.1 Unsourced multiple access

We follow the framework proposed by Polyanskiy [4] in which communication occurs in blocks of n channel uses (measurements) and the task of a receiver is to correctly identify K messages, each consisting of B bits. A typical regime of interest is $n = 30000$, $K = 300$ and $B = 100$. As in [4], we model this scenario using a Gaussian multiple access channel in which the receiver observes

$$y = \sum_{i=1}^{\mathcal{C}} s_i x_i + z,$$

where $\mathcal{C} := 2^B$ is the total number of possible B -bit messages, $s = [s_1 \ \dots \ s_{\mathcal{C}}]^T \in \{0, 1\}^{\mathcal{C}}$ is the activity pattern with Hamming weight K , $x_i \in \mathbb{R}^n$ is the codeword corresponding to message i and $z \sim \mathcal{N}(0, I_n)$ is additive white Gaussian noise (AWGN). Writing $X = [x_1 \ \dots \ x_{\mathcal{C}}]$, we may represent the encoding as the matrix equation

$$y = Xs + z.$$

One popular approach to the problem is slotted ALOHA [5] in which a frame is divided into subframes and each message is sent over a randomly selected subframe (or slot). Decoding of a given message is successful if it encounters no collision with other messages. This no-collision requirement imposes a severe constraint on the number of messages that can be sent, which can be partially alleviated using Coded Slotted ALOHA [6, 7, 8] in which messages are sent to multiple (two or more) slots and information is passed between slots to recover messages originally lost to collisions. These contributions assume synchronized transmission and perfect interference cancellation. The asynchronous model is considered in [9] under the assumption of perfect interference cancellation.

The point is made in [10] that slotted ALOHA only supports single user detection within each slot, and the authors instead propose T -fold ALOHA in which up to T messages can be jointly decoded in each slot. Within each slot, messages are encoded using a product of two codes: the inner code enables the receiver to decode the modulo-2 sum of all codewords transmitted and the outer code allows the receiver to recover the individual messages from the sum. Rather than proposing a specific algorithm, the authors obtain theoretical bounds on performance using fundamental coding limits for the inner code. These bounds are shown to give an improvement upon standard slotted ALOHA-based methods in terms of energy per bit. In practice, a suitable coding and decoding scheme for the binary memoryless channel must be chosen (for example LDPC codes and message-passing recovery), and the choice of decoding algorithm involves a tradeoff between computational efficiency and suboptimality with respect to the coding limits.

Further bounds are given in [11] in which the replica symmetric formula is used to analyze the achievable rate of approximate message passing for decoding the inner code, and improved bounds are reported. The unsourced multiple access paradigm was further extended to the case of massive MIMO and multiple antennas in [12].

In [13], a method was proposed which also performs joint decoding within a slotted framework, but with a number of differences to the approach in [10]. Rather than assigning messages to slots at random, the slots in which a message is repeated are a deterministic function of the message itself. Within each slot, part of the message is encoded by an LDPC code. A preamble picks an interleaver (permutation) for the LDPC code, and is encoded using either a random Bernoulli or a BCH codebook. Either one-step thresholding (treat interference as noise) or nonnegative least squares is used to decode the preamble, and this information is then fed into a message passing algorithm for decoding the interleaved LDPC code. Information is also allowed to pass between slots: when a message is decoded, its corresponding codewords are ‘peeled off’ all slots within which it was transmitted. The authors demonstrate further energy per bit improvements over [10].

In [14], a divide-and-conquer compressed sensing approach is taken in which the message is encoded with redundancy in the form of parity check constraints determined by a linear block code. The message is then split into *patches*, each of which is encoded using a BCH codebook. Each patch is reconstructed using nonnegative least squares, and the messages are then stitched together using a tree decoder. The full paper [15] analyzes two different parity bit allocation strategies in the limit as the number of users and their corresponding payloads tend to infinity. The number of channel uses needed and the computational complexity associated with these allocation strategies are explicitly characterized for various scaling regimes. Further improvements in terms of energy per bit are reported, though for a more modest number of bits than in [10, 13].

1.2 Main contributions

Chirps are a broad class of signals with optimal auto- and cross-correlation properties that encourage their use in multiuser sonar and radar, and in communications (for references to these applications see [16, 17]).

We propose CHIRRUP,¹ a novel compressed sensing algorithm for unsourced multiple access which employs *binary* chirp coding and the chirp reconstruction algorithm originally proposed in [1]. More specifically, we enhance the original algorithm in a number of ways which are inspired by the aforementioned prior work. In particular:

1. Following [5, 6, 7, 10, 13], we incorporate the algorithm into a slotting framework in which each message is sent to precisely two slots. It is this repeated transmission of chirps which inspires the name of the algorithm. As in [13], the slotting pattern is a deterministic function of the message bits, which allows information to be encoded in the slotting pattern. We design a coding scheme which allows messages to be passed between slots: Each message is assigned two slots, and when a message is decoded in one slot the corresponding codewords in its twin slot is peeled off, in a similar spirit to [13]. Our algorithm cycles through the slots allowing the decoded results to propagate.
2. We build in the option to split the message into a small number of patches with the addition of extra parity check constraints. Each patch is decoded separately, and the tree decoder proposed in [14] is then employed to patch together the full message.

¹MATLAB code is available to download from <https://github.com/ajthompson42/CHIRRUP>.

One significant advantage of our approach is computational efficiency. One additional contribution of this paper is to analyse the complexity of the chirp reconstruction algorithm, and show that its worst-case complexity is $\mathcal{O}[nK(K + \log_2^2 n)]$, where n is the codebook length and K is the number of transmitted messages. In contrast to almost all algorithms for compressed sensing, our enhanced chirp reconstruction algorithm is sublinear in the number of codewords, which makes it an attractive choice for a problem whose scale is massive compared to those that are typically considered. We present numerical results which show that the performance of our algorithm in terms of energy per bit is comparable with the results presented in [10, 13, 14] and the maximum number of messages that can be recovered is within a factor of 2 of the results in [10, 13, 14]. Meanwhile we present for the first time in this context information on practical computational efficiency in the form of running time results. Our results therefore represent a benchmark against which other practical algorithms for unsourced multiple access may be compared.

The structure of the rest of the paper is as follows. In Section 2, we describe the original chirp reconstruction algorithm proposed in [1] and give a worst-case complexity result. In Section 3, we describe in detail the enhancements to the original algorithm which lead to improved performance on the unsourced multiple access problem. In Section 4, we present the results of our numerical experimentation, before offering conclusions and reflections on future work in Section 6.

2 The chirp reconstruction algorithm

Our basic approach is to assign to each message a codeword from a codebook of binary chirps. Assuming that $n = 2^m$ for some m , we index the 2^m entries by a binary m -tuple $v \in \mathbb{Z}_2^m$ and consider *binary chirp* codewords of the form $\{\phi_{P,b}\}$ where

$$\{\phi_{P,b}\}_v := \sqrt{Q} \cdot i^{2b^T v + v^T P v}, \quad (1)$$

where addition in the exponent is modulo 4.² The codebook consists of all $\{\phi_{P,b}\}$ where $b \in \mathbb{Z}_2^m$ is a binary m -tuple and $P \in \mathbb{Z}_2^{m \times m}$ is an $m \times m$ binary symmetric matrix. A binary m -tuple b is uniquely determined by m bits, while an $m \times m$ binary symmetric matrix P is determined by $\frac{1}{2}m(m+1)$ bits, which means that the full codebook encodes $m + \frac{1}{2}m(m+1) = \frac{1}{2}m(m+3)$ bits. Note that, since P is symmetric,

$$v^T P v = \sum_{i,j} v_i P_{ij} v_j = 2 \sum_{i < j} v_i P_{ij} v_j + \sum_i v_i P_{ii}.$$

If a real codebook is desired, we use only the subset of $m \times m$ binary symmetric matrices with zero diagonal, and in this case the codebook encodes slightly fewer bits, namely $m + \frac{1}{2}m(m-1) = \frac{1}{2}m(m+1)$. In either case, note that the number of bits is quadratic in $m = \log_2 n$. Each codeword satisfies $\|\phi_{P,b}\|_2^2 = nQ$, and so the constant Q in (1) represents the transmitted power per message. Given K active messages, our measurements therefore take the form

$$y = \sum_{l=1}^K \phi_{P_l, b_l} + z, \quad (2)$$

²All quantities in the exponent are defined over \mathbb{Z}_2 , but the exponent itself is computed over \mathbb{Z}_4 , for example

$$2 \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = 2 + 3 \equiv 1 \pmod{4}.$$

where $z \sim \mathcal{N}(0, I_n)$.

The main building block of our decoding strategy is the *chirp reconstruction algorithm* originally proposed by Howard et al. [1]. The complexity of this algorithm is sublinear in the number of codewords, a vital consideration in the problem at hand if the number of codewords is of the order 2^{100} . We will show that the worst-case complexity of the algorithm is $\mathcal{O}[nK(K + \log_2^2 n)]$, where K is the number of messages. The algorithm is greedy in nature, finding in each iteration a single active component (codeword) and performing a least-squares fit over all components so far discovered, and in that sense it is similar to the popular Orthogonal Matching Pursuit (OMP) algorithm. Where it differs from OMP is in the procedure for identifying the new component. OMP calculates the correlation between the current measurement residual and each column and selects the column whose correlation has the largest absolute value. Finding the maximum value from a list of \mathcal{C} candidates in this way is an $\mathcal{O}(\mathcal{C})$ procedure and therefore computationally intractable for the problem at hand. By contrast, the chirp reconstruction algorithm performs a small number of $\mathcal{O}(n \log_2 n)$ transforms to identify the matrix P row-by-row, and subsequently to identify the vector b , corresponding to an active codeword. The chirp reconstruction algorithm is summarized in Algorithm 1. The iteration limit S should be chosen to be a small multiple of the number of messages.³

Algorithm 1 Chirp Reconstruction Algorithm [1]

Inputs: Measurements $y \in \mathbb{R}^n$ where $n = 2^m$; target sparsity S .

Initializations: $\Phi = [\]$; iteration count $s = 0$.

while $\|y\| > \epsilon$ and $s \leq S$ **do**

$s \leftarrow s + 1$

$(P_s, b_s) \leftarrow \text{findPb}(y)$

$\Phi \leftarrow [\Phi \ \phi_{P_s, b_s}]$

$c_s \leftarrow \Phi^\dagger y$

$y \leftarrow y - \Phi c_s$

end while

Output: $\{P_s, b_s, c_s\}$.

findPb denotes the subroutine for identifying a new component, which is based upon the application of shift-and-multiply to the measurements. More specifically, we shift the binary indices of y through addition of some $e \in \mathbb{Z}_2^m \setminus 0$, and then multiply componentwise with the complex conjugate of y to obtain a vector $f^e \in \mathbb{Z}_2^m$ whose components are $f_a^e := \overline{y_a} y_{a+e}$. Writing y_a for the component of y indexed by $a \in \mathbb{Z}_2^m$, we have the following expression for f^e .

Proposition 1 For any $e \in \mathbb{Z}_2^m$, the following holds for all $a \in \mathbb{Z}_2^m$.

$$\begin{aligned}
 f_a^e = & \sqrt{Q} \sum_{k=1}^K \{\phi_{P_k, b_k}\}_e (-1)^{e^T P_k a} + \sum_{k \neq l} \{\phi_{P_l, b_l}\}_e \{\phi_{P_l - P_k, b_k - b_l + P_l e}\}_a \\
 & + \overline{z_a} \left[\sum_{k=1}^K \phi_{P_k, b_k} \right]_{a+e} + z_{a+e} \overline{\left[\sum_{k=1}^K \phi_{P_k, b_k} \right]_a} + \overline{z_a} z_{a+e}.
 \end{aligned} \tag{3}$$

³In practice we take S to be three times the expected number of messages.

A proof of Proposition 1 can be found in Appendix A. The right-hand side of (3) is a linear combination of Walsh functions (the first term) plus a linear combination of chirps (the remaining terms). We may think of chirps as being distributed across all Walsh functions to equal degree, and therefore these cross terms appear as a uniform noise floor. The Walsh-Hadamard transform $\mathcal{H}(z)$ is the decomposition of a signal $z \in \mathbb{Z}_2^m$ into a basis of Walsh functions, and given $v \in \mathbb{Z}_2^m$ we define its coefficients in the usual way as

$$\{\mathcal{H}(z)\}_v := \frac{1}{2^{m/2}} \sum_{a \in \mathbb{Z}_2^m} (-1)^{v^T a} z_a.$$

The Walsh-Hadamard Transform is the binary analogue of the Discrete Fourier Transform, and applying the Walsh-Hadamard transform \mathcal{H} to (3) results in peaks at ‘frequencies’ $P_k e$ for each P_k in the data. Useful information can thus be obtained by choosing any $e \in \mathbb{Z}_2^m \setminus 0$. In particular, if $e := e_r$ is the r th canonical basis vector in \mathbb{Z}_2^m , peaks will correspond to the r th row (or column) of the $\{P_k\}$ matrices. In this way, an active P_k matrix can be discovered row-by-row. To mitigate cancellation effects for $r > 1$, a second choice of $e := e_r + e_{r-1}$ is also used, the absolute values of the two outputs are summed, and the peak then chosen (see [1] for further explanation). By way of illustration, Figure 1 plots the absolute values of f in the determination of the second column of the P matrix of a component. The figure illustrates how components corresponding to chirps present in the signal are prominent above the ‘noise floor’. Having found P_k , the corresponding b_k can be determined by *dechirping* the signal by calculating $i^{-a^T P_k a} y_a$ for each $a \in \mathbb{Z}_2^m$ and applying the Walsh-Hadamard transform \mathcal{H} to the result. The **findPb** algorithm is summarized in Algorithm 2.

Algorithm 2 findPb [1]

Inputs: Residual $y \in \mathbb{R}^n$.

Initializations: $P = []$.

```

for  $r = 1$  to  $m$  do
   $f_a \leftarrow \overline{y_a} y_{a+e_r}$  for  $a \in \mathbb{Z}_2^m$ 
  if  $r=1$  then
     $h \leftarrow |\mathcal{H}(f)|$ 
  else
     $g_a \leftarrow \overline{y_a} y_{a+e_r+e_{r-1}}$  for  $a \in \mathbb{Z}_2^m$ 
     $h \leftarrow |\mathcal{H}(f)| + |\mathcal{H}(g)|$ 
  end if
   $p_r \leftarrow$  location of largest entry of  $h$ 
end for
 $P \leftarrow [p_1 \ \dots \ p_m]$ 
 $z_a \leftarrow i^{-a^T P a} y_a$  for  $a \in \mathbb{Z}_2^m$ 
 $w \leftarrow |\mathcal{H}(z)|$ 
 $b \leftarrow$  location of largest entry of  $w$ 

```

Outputs: P, b .

The most expensive operations in each iteration of Algorithm 2 are the Walsh-Hadamard transforms (WHTs), which can be computed in $\mathcal{O}(n \log_2 n)$ operations. Two WHTs must be computed in each iteration to select P , and then a single WHT is required to select b . Since there are $m = \log_2 n$ iterations, the overall complexity of Algorithm 2 is therefore $\mathcal{O}(n \log_2^2 n)$. As well as a

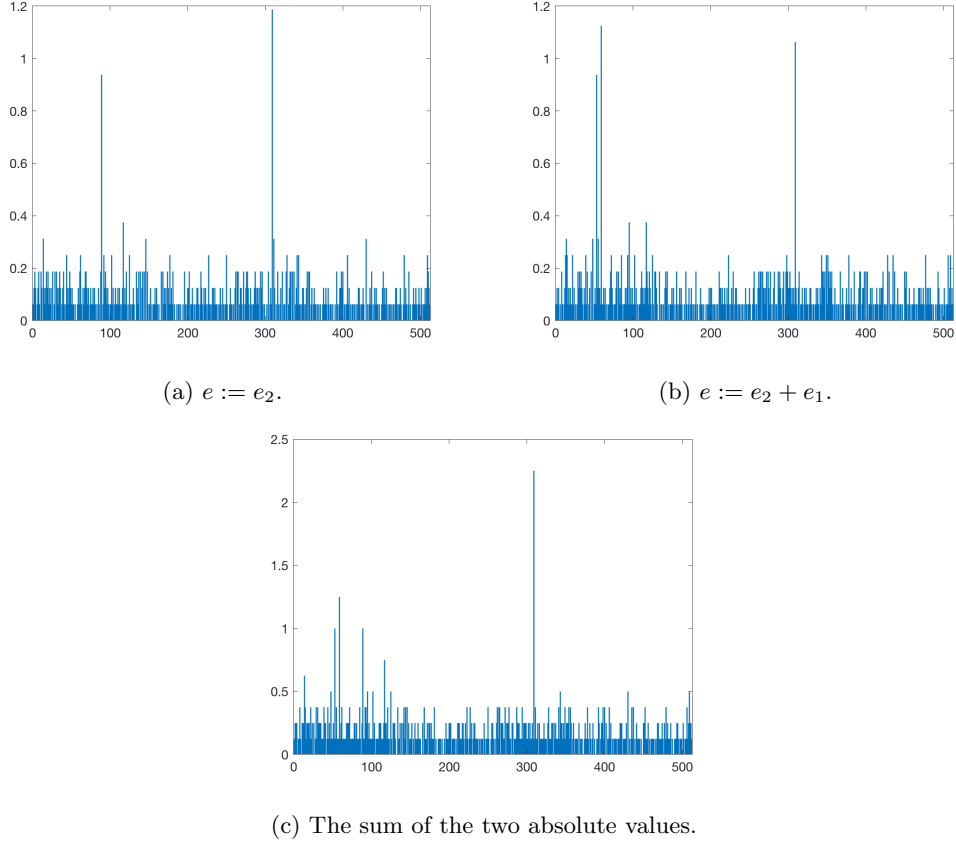


Figure 1: A measurement consisting of the noiseless superposition of 4 chirp components ($m = 10$) is generated, and the figure plots the absolute values of f^e in the determination of the second row of the P matrix in the first iteration. Out of the 512 remaining candidates, a small number emerge above the noise floor. When the absolute f^e scores for $e := e_2$ and $e := e_2 + e_1$ are summed, a single component emerges, which corresponds to the correct second row.

call to Algorithm 2, each iteration of Algorithm 1 also requires the updating of the least-squares fit to the residual. Rather than performing these calculations explicitly, a more computationally efficient approach is the one given in [18], in which a reduced QR factorization is updated, which requires $\mathcal{O}(nK)$ operations per iteration. Since there are $\mathcal{O}(K)$ iterations in Algorithm 1, its overall complexity is therefore $\mathcal{O}[nK(K + \log_2^2 n)]$. Note that, since $n \ll \mathcal{C}$, the algorithm is massively sublinear in \mathcal{C} .

We also consider an enhancement to Algorithm 2 which gives an improvement in performance at the cost of a higher worst-case complexity. The change is to allow the algorithm to test a number of candidates c for each row of the matrix P corresponding to the largest c entries of h (see Algorithm 2). The choice of P corresponds to selecting a branch of a tree in which the r th level of the tree represents the choice of the r th row of P , and in which each node has c children. The

tree is explored in a depth-first fashion, starting with the candidates at each level for which h is greatest as in Algorithm 2. When a branch of the tree is reached, the choice of (P, b) is accepted if the absolute value of w (see Algorithm 2) exceeds some multiple α of the root mean squared error of w restricted to the other components: we find $\alpha := 3$ to be a sensible choice. If not, we retrace our steps back along the branch of the tree, trying the other candidates for the m th row, followed by the $(m - 1)$ th row, and so on. If the whole tree is explored and no solution is found that satisfies the acceptance criterion, a random (P, b) is generated.

Analyzing the worst-case complexity of this enhanced Algorithm 2, we see that, if the entire tree is explored, $2 \times c^r$ WHTs must be computed at level r of the tree to select P , the sum total of which may be bounded above by $2c^{m+1}$, and c^m WHTs must be computed at each of the branches. The total number of WHT computations is therefore bounded above by $3c^{m+1} = 3cn^{\log_2 c}$. The worst-case complexity of the enhanced Algorithm 2 is therefore $\mathcal{O}(n^{1+\log_2 c} \log_2 n)$. For example, we have $\mathcal{O}(n^2 \log_2 n)$ when $c = 2$ and $\mathcal{O}(n^3 \log_2 n)$ when $c = 4$. We find that $c = 3$ is quite sufficient in practice. We also observe that in practice running times only increase by a linear factor compared to the basic algorithm since it is rare for the entire tree to be searched. The worst-case complexity of the full chirp reconstruction algorithm, Algorithm 1, with the enhancement, is therefore $\mathcal{O}[nK(K + n^{1+\log_2 c} \log_2 n)]$, which is still sublinear in \mathcal{C} since $c \ll n$.

3 The CHIRRUP algorithm

The number of messages that can be successfully decoding using the basic algorithm scales somewhat poorly with codeword length. We thus consider three enhancements to the basic decoder: *slotting*, *message passing* and *bit partitioning*. In this section, we assume the use of complex binary chirps, but each of the enhancements can also be used in the case of real binary chirps. In Section 4, we will present numerical results using both complex and real binary chirps. We refer to the algorithm obtained by incorporating each of these enhancements into Algorithm 1 as CHIRRUP.

3.1 Slotting

An enhancement which allows a greater number of messages to be recovered is slotting, in which a codeword length $n = 2^{m+p}$ is divided into 2^p slots, with each message being sent to one or more of the slots. Equivalently, each message is assigned a sparse codeword which is zero except for the blocks to which the message is sent. The chirp reconstruction algorithm is then used to identify the messages within each slot, and the results are combined. Slotting can also be leveraged to send information by allowing some of the message bits to encode the slot location.

Slotting is frequently used in multiple access as a tool to manage interference, for example in slotted ALOHA protocols [5]. Slotting schemes have recently been proposed in the context of unsourced multiple access in [4, 10, 13]. If the same ‘inner’ code is used in each slot, this approach can also be viewed as a special case of a product code where the ‘outer’ codebook matrix is simply the identity matrix.

Since slots must be assigned to messages *a priori*, messages must be slotted in an uncoordinated fashion, which means that it is not possible to ensure a precise distribution of messages between the slots. However, if the slots are equally distributed between the messages and if the transmitted messages are independent of this distribution, one expects the number of active messages over the slots to follow a multinomial distribution with equal probabilities per slot.

3.2 Message passing and a peeling decoder

Message passing algorithms are often used for solving inference problems on graphs, and in the context of wireless communication belief propagation decoders are frequently used for codes built from sparse graphs, such as LDPC codes and Turbo codes. For example in LDPC codes, the individual parity checks that comprise the code are decoded separately and the information shared iteratively. Exporting the principle to slotted chirp reconstruction, one might expect it to be advantageous to assign each message to multiple slots and to allow the slots to propagate the decoded messages in some iterative fashion.

Our approach in general is as follows. The bits used to encode (P, b) are given a dual function, namely also to encode the q slots to which each message is sent. The decoding algorithm makes a number of passes (cycles) through all the slots, and whenever a (P, b) pair is found, the slots to which it was sent can be identified and the corresponding chirp component ‘peeled off’ from each set of measurements. In this fashion, the information decoded in each slot is propagated.

We found that sending each message to *exactly two* ($q = 2$) of the 2^p slots was optimal in practice. In this context, we also found a way to engineer the encoding of additional bits in the slot pattern, as follows. We transmit $\frac{1}{2}m(m+3) + p - 1$ bits by assigning each message to $q = 2$ out of the 2^p slots. Assuming $p \leq \frac{1}{2}m(m+3) - 1$, we append p bits to encode a *primary slot location*. From the bits used to encode (P, b) , we use an arbitrary subset of size p to encode a *translate*. By a translate, we mean that it gives the secondary slot location when it is added (using binary addition) to the primary slot location.⁴ To allow the primary and secondary slots to be distinguished, we fix a single check digit in the P matrix to be 0 for the primary slot and 1 for the secondary slot, deducting 1 from the total number of bits transmitted.

Our decoding algorithm makes d cycles through all the slots, in each of which all of the slots are separately decoded.⁵ Whenever a (P, b) pair is found, the check digit in the P matrix reveals whether the current slot is primary or secondary for the corresponding message. Meanwhile the translate implicitly encoded in the given (P, b) determines the other slot to which the message was sent. Regardless of whether the slot is primary or secondary, the primary slot is now known and the corresponding message can be decoded. Provided its corresponding coefficient c is suitably close to 1,⁶ the check digit is flipped and the resulting (P, b) is passed to the other slot. Each time decoding is attempted in a given slot, a list of (P, b) already known to be active in that slot (either from its own previous decodings or from messages passed from other slots) is compiled. The corresponding components are then ‘peeled off’ from the signal before the iterative procedure begins, allowing more components to be found.

3.3 Bit partitioning

We have already noted that a given binary chirp codebook places an upper limit on the length of messages that can be transmitted, and the use of slotting further constrains the length of messages. Indeed a tradeoff can be observed at the outset here: for a fixed codebook length $n = 2^m$, more slots allow more messages to be transmitted while reducing the length of messages that can be transmitted. This tradeoff will be explored quantitatively in Section 4. One way to allow longer messages to be transmitted is to partition the message into patches in such a way that the patches

⁴Specifically, we first use bits used to encode b , and if needed some of the bits used to encode P .

⁵We find $d = 5$ to be universally sufficient in practice.

⁶Various criteria are possible here and our choices are *ad hoc*. For complex chirps, we found the condition $|c - 1| < 0.3$ to work well. For real chirps, we require $|\operatorname{Re}(c) - 1| < 0.1$ and $|\operatorname{Im}(c)| < 0.1$.

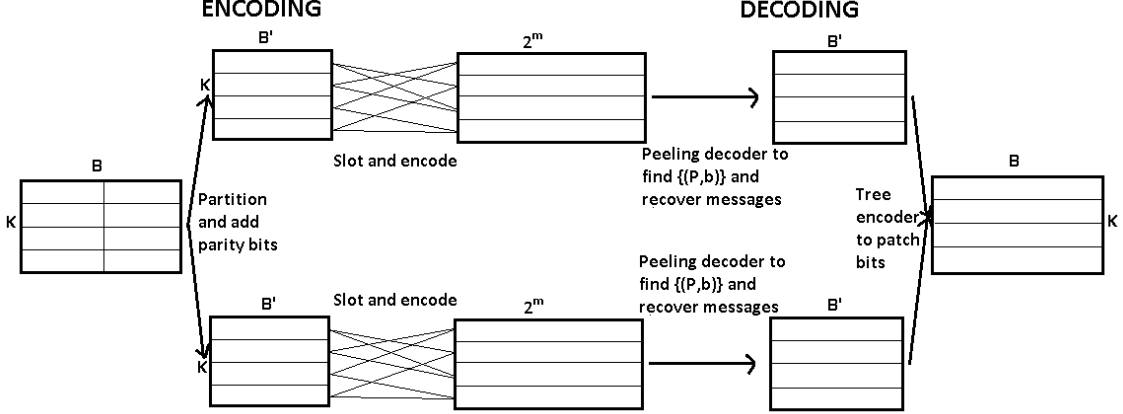


Figure 2: An illustration of the slotting and patching schemes used in the CHIRRUP algorithm. For simplicity, we here illustrate for 2 patches and $K = 4$.

can be decoded separately and then ‘patched together’ after the event. In this regard, we adopt the approach recently proposed in [14] in which each patch is encoded using a linear block code which introduces redundancy in the form of random parity checks. The patches are then decoded separately and ‘patched together’ after the event using a tree encoder. We refer the reader to [14] for further details of the method.

In our case, we choose an integer partition parameter r and partition into 2^r patches ($r = 0$ is a single patch, $r = 1$ means two patches, etc.). We partition the length- $n = 2^{m+p+r}$ codeword into 2^r sub-blocks, each of length 2^{m+p} . We then further split each sub-block into 2^p slots of length 2^m , as described in Section 3.2. In this way a message consisting of $\frac{1}{2}m(m+3) + p - 1$ bits can be sent patchwise in each of the sub-blocks. In keeping with [14], we append l_i random parity check bits to patch i , for $i = 2, \dots, 2^r$ (we never assign parity check bits to the first patch). These parity check bits are needed to patch together the messages, but do not serve the purpose of transmitting information.⁷ The number of bits of information transmitted is therefore

$$B = 2^r \left\{ \frac{1}{2}m(m+3) + p - 1 \right\} - \sum_{i=2}^r l_i.$$

An illustration of the slotting and patching scheme used in the CHIRRUP algorithm is shown in Figure 2.

Concerning complexity, CHIRRUP makes $d \cdot 2^{p+r}$ calls to the chirp reconstruction algorithm, where d is a small constant. Meanwhile the complexity of the chirp reconstruction algorithm was shown in Section 2 to be $\mathcal{O}[\tilde{n}K(K + \log_2^2 \tilde{n})]$, where $\tilde{n} := 2^m$. Combining these two observations, it follows that an analogous complexity result holds for CHIRRUP as for the chirp reconstruction algorithm, that is $\mathcal{O}[nK(K + \log_2^2 n)]$.

⁷The choice of $\{l_i\}$ depends upon the number of messages being transmitted (see [14] for an illuminating theoretical analysis).

4 Numerical experiments

We wish to assess the ability of the CHIRRAP algorithm described in Sections 2 and 3 to decode a given number of messages. We choose to follow the framework proposed by Polyanskiy in [4] in which the goal is to output a list of which messages were sent. Writing $\Lambda \subseteq \{1, \dots, \mathcal{C}\}$ for the set of transmitted messages, so that $|\Lambda| = K$, we seek an algorithm which outputs $\Gamma \subseteq \{1, \dots, \mathcal{C}\}$ such that $|\Gamma| = K$.⁸ Given $i \in \Lambda$, we write E_i for the error event associated with message i , namely that $i \notin \Gamma$. In keeping with [4], we choose to assume an average-case model where the entries in Λ are distributed uniformly at random on $\{1, \dots, \mathcal{C}\}$. We consider decoding to be successful if the *per-user* probability of failure

$$\mathcal{P} := \frac{1}{K} \sum_{i \in \Lambda} \mathbb{P}(E_i)$$

is below some threshold ϵ , where $\epsilon > 0$ is a parameter to be chosen. An approach that has become popular [4, 10, 13, 14] is then to study the trade-off between the transmission power Q and the number of transmitted messages K , for a given failure probability threshold. One way to view transmission power is in terms of the average energy per bit required to transmit each message, denoted by E_b/N_0 , and defined to be

$$\frac{E_b}{N_0} := \frac{nQ}{2B}.$$

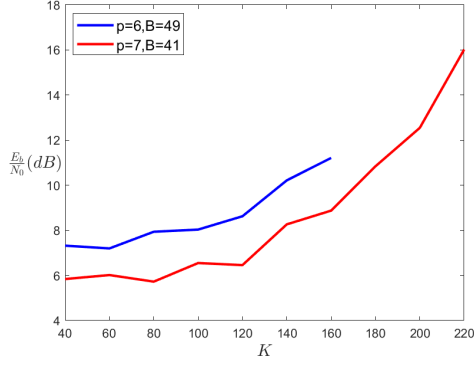
We plot minimum required E_b/N_0 against K for our enhanced chirp reconstruction to achieve a success probability of above 0.95 ($\epsilon = 0.05$). We fix the code length⁹ to be $n = 2^{14} = 16,384$. Various slotting and patching configurations as described in Section 3 are possible for the CHIRRAP algorithm: we vary the number of slots 2^p and number of patches 2^r for different positive integers p and r . We restrict to parameter choices for which the number of bits exceeds 40, for which the number of permissible messages exceeds 80, and for which the algorithm running time is not prohibitive. We find that, generally speaking, a choice of $p \in \{5, 6, 7\}$ is required to meet these conditions: larger p leads to too short or too few messages, while smaller p leads to too few messages or prohibitive run time.¹⁰ We also need to choose the number of parity check bits when the number of patches exceeds 1: we find by experimentation that $l = [0 \ 15]$ and $l = [0 \ 10 \ 10 \ 15]$ are sensible choices for $r = 1$ and $r = 2$ respectively. These choices are made with a view to minimizing the number of bits devoted to parity checking whilst keeping the probability of information loss in the patching process low. We find that only a modest number of patches is beneficial (up to 4), which deviates somewhat from the implementation in [14] in which 11 patches were used. It should be noted that the much more computationally demanding nonnegative least squares is used as the reconstruction algorithm in [14], which necessitates the splitting up of messages into more patches to make the reconstructions computationally tractable.

Plots of E_b/N_0 (displayed on a decibel scale) against K are shown for 1, 2 and 4 patches in Figures 3a, 3c and 3e respectively. We observe that the slotting parameter p allows for a trade-off between both number of transmitted messages and energy per bit (E_b/N_0) on the one hand and

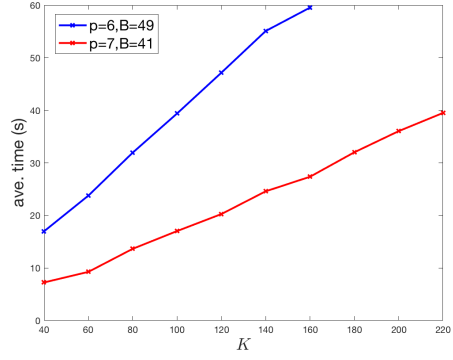
⁸ In our experiments we follow the framework in [4] where it is assumed that the algorithm knows the number of messages K , which may not be practically realistic. It is worth noting, however, that our algorithm does not need to know the number of messages, and indeed its performance improves if we allow it to find more than K messages.

⁹ This corresponds to a real codeword length of $2^{15} = 32,768$.

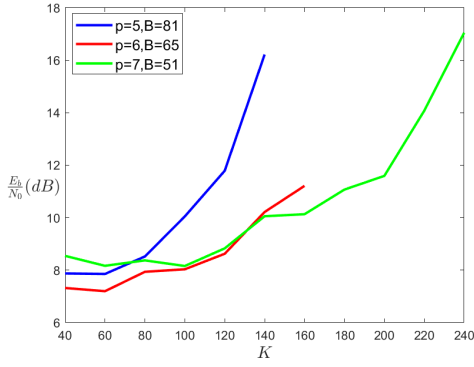
¹⁰ Since the number of messages decodable by the chirp reconstruction algorithm grows sublinearly with codebook length, when p is small the number of messages decreases. On the other hand, if p is taken too large and the codebook length is very small, the chirp reconstruction algorithm ceases to perform well.



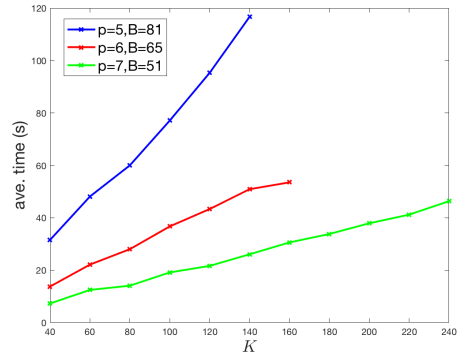
(a) E_b/N_0 for 1 patch.



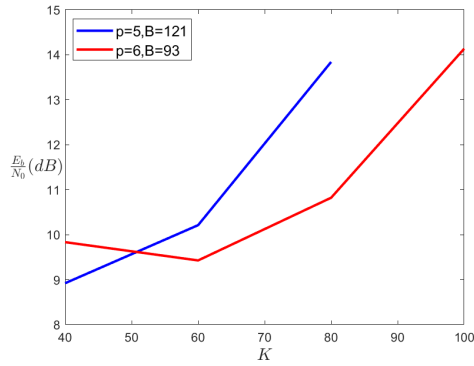
(b) Running time for 1 patch.



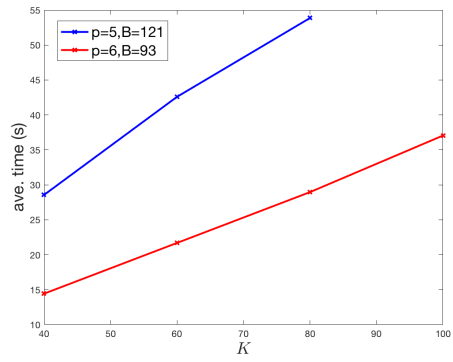
(c) E_b/N_0 for 2 patches.



(d) Running time for 2 patches.



(e) E_b/N_0 for 4 patches.



(f) Running time for 4 patches.

Figure 3: Energy per bit (subfigures (a), (c) and (e)) and average computation time (subfigures (b), (d) and (f)) for CHIRUP with complex chirps.

message length (B) on the other hand. As the number of slots is increased, more messages can be decoded and E_b/N_0 generally decreases for a given K . This gain is achieved at the cost of reducing the codeword size within each slot, which decreases the message length.

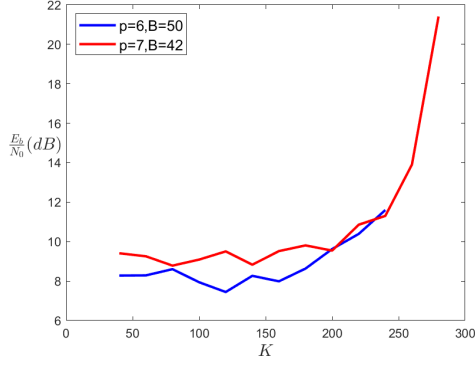
We also observe that increasing the number of patches, while keeping the number of slots fixed, both increases the message length and decreases the number of messages that can be decoded, thus representing a second trade-off. As a rough guide, 1 patch appears to be preferable for modest bit lengths of under 50 bits, 2 patches are preferable for between 50 and 80 bits and 4 patches are preferable for upwards of 80 bits.

We observe in both Figures 3c and 3e that the curves for $p = 5$ and $p = 6$ cross each other, which might appear surprising. The authors believe that the explanation for this behaviour lies in suboptimal tuning of a certain parameter, namely how ‘close’ in value to 1 a coefficient must be for its corresponding chirp component to be retained (see Section 3.2). One of the advantages of CHIRUP is that no tuning of parameters is required: the same parameter choices were used throughout all the tests. However, our observation is that the threshold selected is somewhat suboptimal for the case of a small number of messages (e.g. $K = 40$), and this accounts for the unusual crossing behaviour.

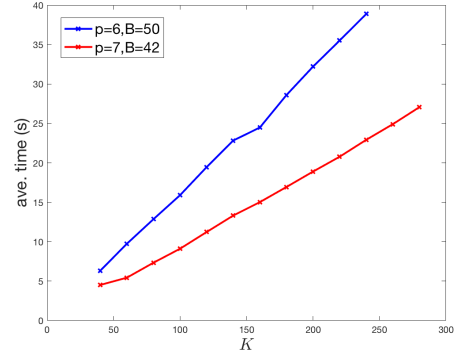
Average computation time is displayed for 1, 2 and 4 patches in Figures 3b, 3d and 3f respectively. For a given number of messages K , the energy per bit E_b/N_0 is set to the smallest which permits a success probability of 0.95 as determined by our experiments. The computations were carried out on a 2.2 GHz MacBook Air. We observe that the CHIRUP algorithm can be run in a few seconds with desktop computing power. We observe a tradeoff between message length and running time: longer messages require larger binary chirp codebooks, which makes the algorithm more computationally burdensome.

Real codebooks. We noted that a codelength of $n = 2^{14}$ corresponds to a real codelength of 2^{15} . Rather than using complex binary chirps, an alternative approach is to restrict to real binary chirps from the outset. In this case, a fair comparison is to take $n = 2^{15}$. We performed analogous numerical experiments for real codebooks as for complex codebooks. Plots of energy per bit against number of messages are displayed for 1, 2 and 4 patches in Figures 4a, 4c and 4e respectively, and plots of running time against number of messages for 1, 2 and 4 patches are displayed in Figures 3b, 3d and 3f respectively.

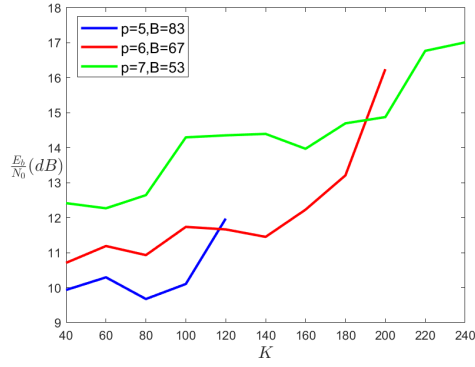
We observe a somewhat close correspondence between the results for complex and real binary chirps, which reflects the fact that the corresponding codebooks are geometrically equivalent. However, for some parameter choices real chirps allow for a better tradeoff between number of bits and number of messages, especially for $r = 0$, and to some extent for $r = 1$. We believe that this improvement is not due to a fundamental difference between complex and real binary chirp codebooks, but rather to a different condition for deciding whether the coefficient corresponding to a retrieved component is close to 1. The results therefore emphasize how sensitive the algorithm is to the choice of this condition, and that it certainly pays to make use of the prior knowledge that all signal coefficients are equal to 1. A thorough investigation of how to optimize the condition is unfortunately a computationally intimidating task, which prevented further exploration of this behaviour.



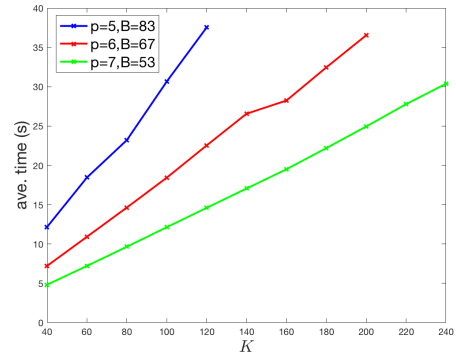
(a) E_b/N_0 for 1 patch.



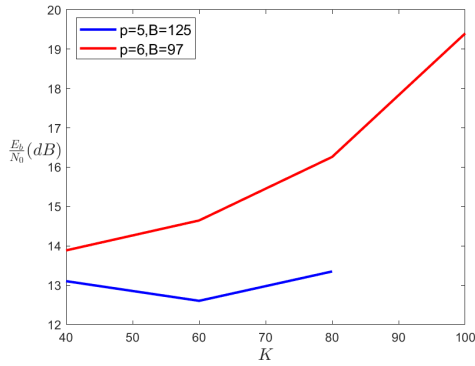
(b) Running time for 1 patch.



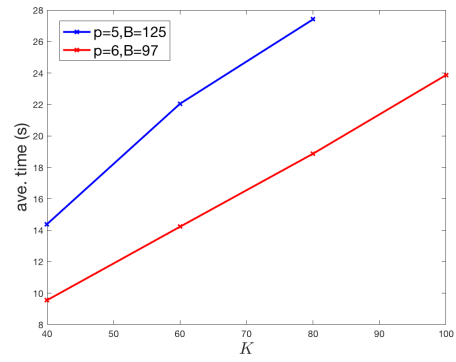
(c) E_b/N_0 for 2 patches.



(d) Running time for 2 patches.



(e) E_b/N_0 for 4 patches.



(f) Running time for 4 patches.

Figure 4: Energy per bit (subfigures (a), (c) and (e)) and average computation time (subfigures (b), (d) and (f)) for CHIRUP with real chirps.

5 Theoretical comparison: One Step Thresholding

Among the plethora of algorithms for compressed sensing reconstruction, the CHIRRUP algorithm is one of the few with the distinction of having running time sublinear in the signal dimension. It is nonetheless interesting to know how the performance of CHIRRUP would compare to other algorithms which scale poorly in this context. In this regard, we choose the very simple One Step Thresholding (OST) algorithm [19], which is also known as *treat interference as noise* (TIN) [4]. The algorithm¹¹ is stated succinctly in Algorithm 3. Assuming that the entries of the codebook

Algorithm 3 One Step Thresholding [19]

Inputs: $y \in \mathbb{C}^n$, $X \in \mathbb{C}^{n \times \mathcal{C}}$, K .

1. $g = X^*y$.
2. $\Gamma := \{i \text{ corresponding to the } K \text{ largest } g_i\}$.

Outputs: Γ .

matrix X are i.i.d. zero-mean Gaussian, it can be shown that, as problem size grows, the entries of g themselves converge in distribution to a Gaussian distribution, with nonzero mean for the entries corresponding to transmitted messages and zero mean for the other entries. The precise behaviour of OST can thus be determined in the asymptotic limit. Writing

$$\delta = \lim_{K \rightarrow \infty} \frac{K}{\mathcal{C}} \quad \text{and} \quad \rho = \lim_{K \rightarrow \infty} \frac{K + 1/Q}{n},$$

we find that, for a given failure probability ϵ (see Section 4), there exists a phase transition threshold in terms of δ and ρ . Note the striking interpretation: for a given problem size, OST is subject to a fundamental limit on the sum of the number of messages and the inverse of the power. This quantity $K + 1/Q$ can be viewed as an ‘effective number of messages’, which is a combination of the true number of messages and the equivocation arising from the noise. The asymptotic behaviour of OST for random coding has been previously derived in [20] using an information-theoretic approach and quantified in [4, 10]. In Appendix B, we provide an alternative derivation which highlights the “inverse of signal power = effective sparsity” interpretation.

Moreover, it has been observed that the average-case behaviour of many compressed sensing algorithms remains indistinguishable when Gaussian random codebooks are replaced with many other deterministic codebooks, including Delsarte-Goethals frames which are built from binary chirps [21]. These asymptotic results for Gaussian matrices would therefore be expected to also accurately predict the performance of OST with binary chirp codebooks.

We choose the same problem dimensions as in our numerical tests: $n = 2^{15} = 32,768$ and error probability $\epsilon = 0.05$. Using the Gaussian asymptotic approximation, we obtain predictions for the performance of OST for varying numbers of bits ($B = 50$, $B = 75$ and $B = 100$) in Figure 5. We observe that the observed performance of CHIRRUP is comparable to OST, with number of transmitted messages easily within a factor of 2. This is despite the fact that CHIRRUP is sublinear in the signal dimension, in comparison with OST which is superlinear, in fact $\mathcal{O}(\mathcal{C} \log \mathcal{C})$. It

¹¹The OST algorithm described in [19] considers the absolute values of the $\{g_i\}$. We remove the absolute value since we have prior knowledge that all signal coefficients are positive.

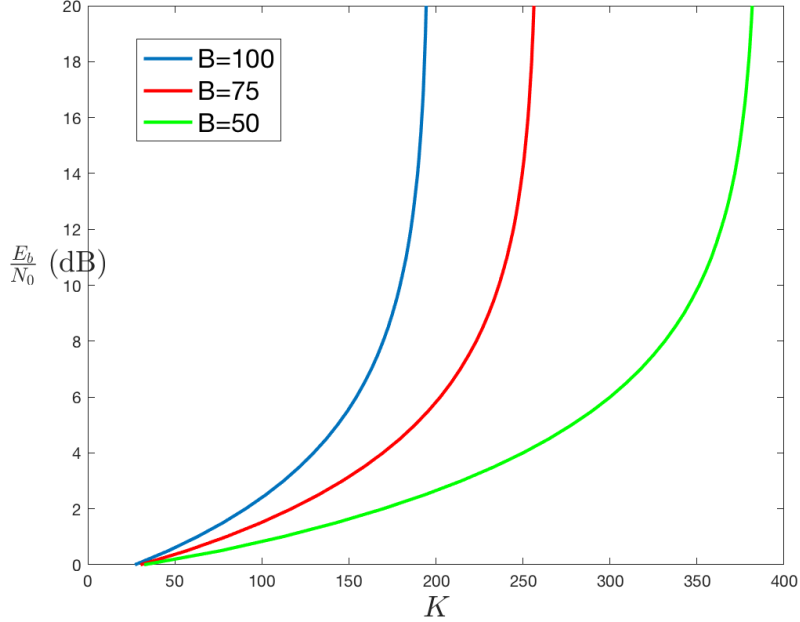


Figure 5: Expected performance of OST for different numbers of messages.

should be emphasized that the plots in Figure 5 are hypothetical: without enormous computational resources the OST algorithm is completely intractable for the problems addressed in this paper.

6 Conclusions and future work

6.1 Conclusions

We have presented CHIRRUP, an algorithm for unsourced multiple access based upon binary chirp coding and the chirp reconstruction algorithm [1]. We have presented the results of numerical simulations, using problem sizes which are realistic for applications, detailing how energy per bit (E_b/N_0) and computing time depends upon number of active users. We believe that this work is the first to engage quantitatively with the performance versus computational efficiency tradeoff for this problem. By contrast, prior contributions either stop short of proposing an actual algorithm (making use of optimal coding bounds for finite blocklength codes) [10, 13] or propose an algorithm without quantifying computational efficiency [14]. We suggest the use of CHIRRUP as a benchmark against which other practical algorithms can be compared numerically. We have made available MATLAB code for CHIRRUP at <https://github.com/ajthompson42/CHIRRUP>.

Viewed another way, CHIRRUP can also be viewed as a combined deterministic measurement scheme and reconstruction algorithm for compressed sensing at massive scale. We have pointed out that CHIRRUP is sublinear in the signal dimension in contrast to almost all other compressed sensing algorithms. For example, not one of the algorithms surveyed in a much-cited review of algo-

rithms for sparse linear inverse problems [22] has sublinear complexity. In this respect, CHIRRUP occupies a unique place among compressed sensing algorithms in its ability to scale to problems at massive scale.

6.2 Future work

Extension to witness-averaging. The *witness-averaging* algorithm, presented in [23], is a modified version of the chirp reconstruction algorithm (Algorithm 1). In this approach, a *Reed-Muller Sieve* codebook is used which consists of binary chirps given by all binary symmetric matrices P but without the phase shifts b . Most significantly, instead of summing two shift-and-multiply computations as in the present paper, all possible shifts are computed and summed. This change has the effect of increasing the complexity of the algorithm so that it is no longer sublinear – in fact $\mathcal{O}(\mathcal{C} \log \mathcal{C})$ – while at the same time drastically improving its decoding performance so that $\mathcal{O}(n)$ components can be recovered. It would be interesting to investigate how the introduction of some measure of witness-averaging affects the performance of CHIRRUP on the number of messages versus computational time tradeoff. It is worth noting that such a change would move CHIRRUP closer in spirit to the approach in [14], in which a more computationally demanding but better performing compressed sensing algorithm is employed over a larger number of message patches.

Application to neighbor discovery. In conventional networks, the overhead of neighbor discovery can be amortized over long data transmissions. However, when there is a massive number of devices, and when transmissions are short and bursty, it has been shown that the overhead of neighbor discovery may considerably reduce data throughput [24].

Neighbor discovery can be formulated as multiuser detection [25] and the coupled compressive sensing approach of [14] was recently extended to the problem of asynchronous neighbor discovery in [26]. Since the number of active devices is typically orders of magnitude smaller than the total number of devices, it is natural to apply ideas from compressed sensing. Synchronous transmission is considered in [27], and asynchronous transmission in [28], where the LASSO algorithm is used to detect active devices. Note that the computational complexity of LASSO is polynomial in the total number of devices. Codewords in [28] are obtained by exponentiating codewords in a 2nd order Reed-Muller code. It is natural to explore replacing LASSO by chirp reconstruction, since this would significantly improve efficiency.

The term *random access* is commonly used to describe a setup where a random subset of users in the network communicate with a base station (BS) in an uncoordinated fashion. The on-off random access channel represents an abstraction where users signal activity by transmitting their identity [29]. The primary objective of the BS is to recover the set of active users. Multiuser detection via One Step Thresholding (OST) is considered in [29, 19].

When a wireless device is equipped with half-duplex hardware, it can either transmit or receive signals in a given time slot, but it cannot do both simultaneously. A method of achieving full-duplex communication using half-duplex radios is proposed in [30]. It is called random on-off-division duplex (RODD) because each user is assigned a unique on-off sequence, and during the off slots each user listens to the signals transmitted by the other users.

A key question in this context is the design of a suitable codebook of on-off sequences. It must be amenable to effective and efficient decoding using compressed sensing algorithms and it must be sparse. Luo and Guo [31, 32] proposed using random Bernoulli codebooks along with a

group testing reconstruction algorithm. Subsequently Zhang and Guo proposed codebooks based on second order Reed-Muller codes along with a modified chirp reconstruction algorithm [33].

The present authors proposed sparse Kerdock matrices as codebooks for the neighbor discovery problem [34]. These matrices share the same row space as certain Delsarte-Goethals frames, while at the same time being extremely sparse. Numerical experiments with One Step Thesholding (OST) and Normalized Iterative Hard Thresholding (NIHT) demonstrate that a higher propotion of neighbors are successfully identified using sparse Kerdock matrices than with codebooks obtained by randomly erasing entries in a Reed-Muller codebook (as proposed in [33]). It is natural to consider replacing OST and NIHT by chirp reconstruction.

A Proof of Proposition 1

Recalling (1), we have

$$\begin{aligned}\overline{\{\phi_{P_k, b_k}\}_a} \{\phi_{P_l, b_l}\}_{a+e} &= Q i^{-2b_k^T a - a^T P_k a} i^{2b_l^T (a+e) + (a+e)^T P_l (a+e)} \\ &= Q i^{2b_l^T e + e^T P_l e + 2(b_l - b_k + P_l e)^T a + a^T (P_l - P_k) a} \\ &= \{\phi_{P_l, b_l}\}_e \{\phi_{P_l - P_k, b_l - b_k + P_l e}\}_a.\end{aligned}\quad (4)$$

When $k \neq l$, we have $P_k \neq P_l$, and (4) consists of a constant term (a chirp in e) multiplied by a chirp in a . However, when $k = l$, we have $P_k = P_l$ and (4) reduces to

$$\overline{\{\phi_{P_k, b_k}\}_a} \{\phi_{P_k, b_k}\}_{a+e} = \sqrt{Q} \{\phi_{P_k, b_k}\}_e (-1)^{e^T P_k a}.\quad (5)$$

Substituting (4) and (5) into (2), we obtain

$$\begin{aligned}f_a = \overline{y_a} y_{a+e} &= \overline{\left[\sum_{k=1}^K \phi_{P_k, b_k} + z \right]_a} \left[\sum_{l=1}^K \phi_{P_l, b_l} + z \right]_{a+e} \\ &= \sum_{k=1}^K \overline{\{\phi_{P_k, b_k}\}_a} \{\phi_{P_k, b_k}\}_{a+e} + \sum_{k \neq l} \overline{\{\phi_{P_k, b_k}\}_a} \{\phi_{P_l, b_l}\}_{a+e} \\ &\quad + \overline{z_a} \left[\sum_{k=1}^K \phi_{P_k, b_k} \right]_{a+e} + z_{a+e} \overline{\left[\sum_{k=1}^K \phi_{P_k, b_k} \right]_a} + \overline{z_a} z_{a+e} \\ &= \sqrt{Q} \sum_{k=1}^K \{\phi_{P_k, b_k}\}_e (-1)^{e^T P_k a} + \sum_{k \neq l} \{\phi_{P_l, b_l}\}_e \{\phi_{P_l - P_k, b_l - b_k + P_l e}\}_a \\ &\quad + \overline{z_a} \left[\sum_{k=1}^K \phi_{P_k, b_k} \right]_{a+e} + z_{a+e} \overline{\left[\sum_{k=1}^K \phi_{P_k, b_k} \right]_a} + \overline{z_a} z_{a+e},\end{aligned}$$

as required.

B Gaussian approximation for OST

As in Section 1, we consider measurements of the form

$$y = Xs + z$$

where $s = [s_1 \dots s_C]^T \in \{0, 1\}^C$ is the activity pattern with Hamming weight K and $z \sim \mathcal{N}(0, I_n)$ is additive white Gaussian noise (AWGN). We now assume that the entries of X are distributed i.i.d $N(0, Q)$. We derive the asymptotic distribution of the entries of g , as defined in Algorithm 3, under the given model for X and z . More precisely, we consider a sequence of problems with parameters (K_r, n_r, C_r) for $r = 1, 2, \dots$ such that

$$\frac{K_r}{C_r} \rightarrow \delta, \quad \frac{K_r + 1/Q}{n_r} \rightarrow \rho.$$

Writing X_i for the i th column of X , we note that $\|X_i\|_2^2/n \sim \frac{Q}{n} \chi_2^2$ which converges in distribution exponentially fast to Q by the law of large numbers as $n \rightarrow \infty$, and so Q continues to represent the transmitted power per message (see Section 1). Writing G_i for the random variable corresponding to the entries g_i , we have a central limit result for each of the G_i .

Theorem 1

$$\frac{G_i}{nQ} \xrightarrow{p} \begin{cases} N(1, \rho) & i \in \Lambda \\ N(0, \rho) & i \in \Lambda^C. \end{cases}$$

Proof: Fix (K, n, Q) and suppose $p \in \Lambda^C$ and consider the random variable

$$Y_i^n := \frac{1}{Q\sqrt{n}} X_{ip} \left(\sum_{j \in \Lambda} X_{ij} + z_i \right).$$

Note that the $\{Y_i^n\}$ are independent for $i = 1, \dots, n$, since they involve entries from different rows of X . Furthermore, we have $\mathbb{E}Y_i = 0$ and $\mathbb{E}Y_i^2 = \frac{K+1/Q}{n} < \infty$. Then by the Berry-Esseen theorem the distribution function, $F_n(x)$, of

$$\frac{\sum_{i=1}^n Y_i}{\sqrt{n \mathbb{E}Y_i^2}} = \frac{\sum_{i=1}^n X_{ip} \left(\sum_{j \in \Lambda} X_{ij} + z_i \right)}{nQ \left(\frac{K+1/Q}{n} \right)^{1/2}}$$

satisfies

$$\sup_x |F_n(x) - \Phi(x)| \leq \frac{\mathbb{E}|Y_i|^3}{\sqrt{n} \left(\frac{K+1/Q}{n} \right)^{3/2}}, \quad (6)$$

where $\Phi(x)$ is the distribution function of the standard Normal distribution. It remains to calculate $\mathbb{E}|Y_i|^3$. In this regard, we note that, if $Z \sim N(0, \sigma^2)$, we have

$$\mathbb{E}|Z|^3 = \frac{2\sqrt{2}\sigma^3}{\sqrt{\pi}}.$$

Now, for all $i = 1, \dots, n$,

$$X_{ip} \sim N(0, Q), \quad \text{and} \quad \sum_{j \in \Lambda} X_{ij} + z_i \sim N(0, KQ + 1),$$

and they are mutually independent. We therefore have

$$\begin{aligned}\mathbb{E}|Y_i^n|^3 &= Q^{-3}n^{-3/2}\mathbb{E}|X_{ip}|^3\mathbb{E}\left(\left|\sum_{j\in\Lambda}X_{ij}+z_i\right|^3\right) \\ &= Q^{-3}n^{-3/2}\cdot\frac{2Q^{3/2}\sqrt{2}}{\sqrt{\pi}}\cdot\frac{2\sqrt{2}}{\sqrt{\pi}}(KQ+1)^{3/2} \\ &= \frac{8}{\pi}\left(\frac{K+1/Q}{n}\right)^{3/2},\end{aligned}$$

which may be substituted into (6) to give

$$\sup_x |F_n(x) - \Phi(x)| \leq \frac{8}{\pi\sqrt{n}}. \quad (7)$$

Now suppose $p \in \Lambda$ and consider the random variable

$$Y_i^n := \frac{1}{Q\sqrt{n}} \left[X_{ip} \left(\sum_{j\in\Lambda} X_{ij} + z_i \right) - Q \right].$$

Note that we again have that the $\{Y_i^n\}$ are independent for $i = 1, \dots, n$, and $\mathbb{E}Y_i = 0$ and $\mathbb{E}Y_i^2 = \frac{K+1+1/Q}{n} < \infty$. Then the Berry-Esseen theorem guarantees that the distribution function, $F_n(x)$, of

$$\frac{\sum_{i=1}^n Y_i}{\sqrt{n}\mathbb{E}Y_i^2} = \frac{\sum_{i=1}^n X_{ip} \left(\sum_{j\in\Lambda} X_{ij} + z_i - Q \right)}{nQ \left(\frac{K+1+1/Q}{n} \right)^{1/2}}$$

satisfies

$$\sup_x |F_n(x) - \Phi(x)| \leq \frac{\mathbb{E}|Y_i^n|^3}{\sqrt{n} \left(\frac{K+1+1/Q}{n} \right)^{3/2}}, \quad (8)$$

and it remains to calculate $\mathbb{E}|Y_i^n|^3$. Writing $Y_i^n = U_i^n + V_i^n$ where

$$U_i^n := \frac{1}{Q\sqrt{n}} (X_{ip}^2 - Q), \quad V_i^n := \frac{1}{Q\sqrt{n}} \left[X_{ip} \left(\sum_{j\in\Lambda\setminus p} X_{ij} + z_i \right) \right],$$

we have

$$\mathbb{E}|Y_i^n|^3 = \mathbb{E}|U_i^n + V_i^n|^3 \leq \mathbb{E}[2\max(|U_i^n|, |V_i^n|)]^3 = 8\max(\mathbb{E}|U_i^n|^3, \mathbb{E}|V_i^n|^3). \quad (9)$$

We have

$$U_i^n \sim \frac{1}{n}(\chi_1^2 - 1),$$

where χ_1^2 denotes the chi-squared distribution with one degree of freedom, and calculation yields

$$\mathbb{E}|U_i^n|^3 = \frac{8}{n^4} \left[3 + \frac{3}{\sqrt{2\pi}} + \sqrt{e} \left(1 - 2\operatorname{erf}\frac{1}{\sqrt{2}} \right) \right].$$

Meanwhile, using a similar argument as in the $p \in \Lambda^C$ case yields

$$\mathbb{E}|V_i^n|^3 = \frac{8}{\pi} \left(\frac{K-1+1/Q}{n} \right)^{3/2}.$$

For sufficiently large K and n – for example if $K \geq 2$ and $n \geq 12$ – it is easy to show that $\mathbb{E}|U_i^n|^3 < \mathbb{E}|V_i^n|^3$, which combines with (9) to give

$$\mathbb{E}|Y_i|^3 \leq \frac{64}{\pi} \left(\frac{K-1+1/Q}{n} \right)^{3/2} \leq \frac{64}{\pi} \left(\frac{K+1+1/Q}{n} \right)^{3/2},$$

which in turn combines with (8) to give

$$\sup_x |F_n(x) - \Phi(x)| \leq \frac{64}{\pi\sqrt{n}} \quad (10)$$

and (7) and (10) together establish the result. \square

We next give details of how Theorem 1 can be used to obtain an accurate asymptotic numerical approximation for the performance of OST. Given a threshold λ , let P_λ and Q_λ be the number of $|g_i|$ above λ from Λ and Λ^C respectively, divided by \mathcal{C} . Then Theorem 1 gives the following corollary.

Corollary 1

$$\mathbb{E}P_\lambda \rightarrow \delta \left[1 - \Phi \left(\frac{\lambda-1}{\rho} \right) \right], \quad (11)$$

$$\mathbb{E}Q_\lambda \rightarrow (1-\delta) \left[1 - \Phi \left(\frac{\lambda}{\rho} \right) \right], \quad (12)$$

where Φ denotes the distribution function of the standard Normal distribution.

Proof:

$$\begin{aligned} \mathbb{E}P_\lambda &= \sum_{i \in \Lambda} \frac{1}{\mathcal{C}} \cdot \mathbb{P}(G_i > \lambda) \\ &= \frac{K}{\mathcal{C}} \mathbb{P}(G_{i^*} > \lambda), \end{aligned}$$

where $i^* \in \Lambda$, from which (11) follows by taking limits and invoking Theorem 1. Similarly,

$$\begin{aligned} \mathbb{E}Q_\lambda &= \sum_{i \in \Lambda^C} \frac{1}{\mathcal{C}} \cdot \mathbb{P}(G_i > \lambda) \\ &= \frac{\mathcal{C}-K}{\mathcal{C}} \mathbb{P}(G_{\hat{i}} > \lambda), \end{aligned}$$

where $\hat{i} \in \Lambda^C$, from which (12) follows by taking limits and invoking Theorem 1. \square

Fixing an acceptance probability ϵ , we solve

$$\mathbb{E}P_\lambda = (1-\epsilon)\delta, \quad \mathbb{E}(P_\lambda + Q_\lambda) = \delta,$$

to obtain a solution for λ and ρ . Hence, asymptotically,

$$\rho = c(\delta, \epsilon),$$

where $c(\delta, \epsilon)$ is a constant which depends only upon δ and ϵ . We therefore have, asymptotically, that

$$\frac{K + 1/Q}{n} = c(\delta, \epsilon),$$

which is what was claimed in Section 5.

References

- [1] S. Howard, R. Calderbank, and S. Searle. A fast reconstruction algorithm for deterministic compressive sensing using second order Reed-Muller codes. In *42nd Annual Conference on Information Sciences and Systems*, Princeton, NJ, March 2008.
- [2] F. Boccardi, R. Heath Jr., A. Lozano, T. Marzetta, and P. Popovski. Five disruptive technology directions for 5G. *IEEE Communications Magazine*, 52(2):74–80, 2014.
- [3] G. Durisi, T. Koch, and P. Popovski. Toward massive, ultrareliable, and low-latency wireless communication with short packets. *Proceedings of the IEEE*, 104(9):1711–1726, 2016.
- [4] Y. Polyanskiy. A perspective on massive random access. In *International Symposium on Information Theory*, Aachen, Germany, June 2017.
- [5] L. Roberts. ALOHA packet system with and without slots and capture. *Computer Communications Review*, 5(2):28–42, 1975.
- [6] E. Casini, R. De Gaudenzi, and O. Herrero. Contention resolution diversity slotted ALOHA (CRDSA): An enhanced random access scheme for satellite access packet networks. *IEEE Transactions on Wireless Communication*, 6(4):1408–1419, 2007.
- [7] G. Liva. Graph-based analysis and optimization of contention resolution diversity slotted ALOHA. *IEEE Transactions on Communications*, 59(2):477–487, 2011.
- [8] E. Paolina, G. Liva, and M. Chiani. Coded slotted ALOHA: a graph-based approach for uncoordinated multiple access. *IEEE Transactions on Information Theory*, 61(12):6815–6832, 2015.
- [9] E. Sandgren, A. Graell i Amat, and F. Brännström. On frame synchronous coded slotted ALOHA: Asymptotic, finite length, and delay analysis. *IEEE Transactions on Communications*, 65(2):691–704, 2017.
- [10] O. Ordentlich and Y. Polyanskiy. Low complexity schemes for the random access Gaussian channel. In *International Symposium on Information Theory*, Aachen, Germany, June 2017.
- [11] A. Fengler, P. Jung, and G. Caire. Sparses for unsourced random access. *arXiv:1901.06234*, 2019.
- [12] G. Fengler, A. and Caire, P. Jung, and S. Haghshatshoar. Massive MIMO unsourced random access. *arXiv:1901.00828*, 2019.

- [13] A. Vem, K. Narayanan, J. Cheng, and J. Chamberland. A user-independent serial interference cancellation based coding scheme for the unsourced random access Gaussian channel. In *IEEE Information Theory Workshop*, Kaohsiung, Taiwan, November 2017.
- [14] V. Amalladinne, A. Vem, D. Kumar Soma, K. Narayanan, and J. Chamberland. A coupled compressive sensing scheme for unsourced multiple access. In *International Conference on Acoustics, Speech and Signal Processing*, Calgary, Canada, April 2018.
- [15] V. Amalladinne, A. Vem, D. Kumar Soma, K. Narayanan, and J. Chamberland. A coupled compressive sensing scheme for uncoordinated multiple access. *arXiv:1809.04745*, 2018.
- [16] A. Brodzik. Characterization of Zak space support of the discrete chirp. *IEEE Transactions on Information Theory*, 53(6):2190–2203, 2007.
- [17] A. Brodzik and A. Tolimieri. Bat chirps with good properties: Zak space construction of perfect polyphase sequences. *IEEE Transactions on Information Theory*, 55(4):1804–1814, 2009.
- [18] T. Blumensath and M. Davies. Gradient pursuits. *IEEE Transactions on Signal Processing*, 56(6):2370–2382, 2008.
- [19] W. Bajwa, R. Calderbank, and D. Mixon. Why Gabor frames? two fundamental measures of coherence and their role in model selection. *Journal of Communication Networks*, 12(4):289–307, 2010.
- [20] Y. Polyanskiy, H. Poor, and S. Verd’u. Channel coding rate in the finite blocklength regime. *IEEE Transactions on Information Theory*, 56(5):2307–2359, 2010.
- [21] H. Monajemi, S. Jafarpour, M. Gavish, Stat 330/CME 362 Collaboration, and D. Donoho. Deterministic matrices matching the compressed sensing phase transitions of Gaussian random matrices. *Proceedings of the National Academy of Sciences*, 110(4):1181–1186, 2013.
- [22] J. Tropp and S. Wright. Computational methods for sparse solution of linear inverse problems. *Proceedings of the IEEE*, 98(6):948–958, 2010.
- [23] R. Calderbank, S. Howard, S. Jafarpour, and J. Kent. Sparse approximation and compressed sensing using the Reed-Muller Sieve. Technical Report TR-888-10, Department of Computer Science, Princeton University, 2010.
- [24] X. Chen and D. Guo. Many-access channels: the Gaussian case with random user activities. In *International Symposium on Information Theory*, pages 3127–3131, Honolulu, Hawaii, June 2014.
- [25] D. Angelosante, E. Biglieri, and M. Lops. Neighbor discovery in wireless networks: a multiuser-detection approach. *Physical Communication*, 3(1):28–36, 2010.
- [26] V. Amalladinne, K. Narayanan, J. Chamberland, and D. Guo. Asynchronous neighbor discovery using coupled compressive sensing. In *International Conference on Acoustics, Speech and Signal Processing*, pages 4569–4573, 2019.
- [27] H. Zhu and G. Giannakis. Exploiting sparse user activity in multiuser detection. *IEEE Transactions on Communications*, 59(2):454–465, 2011.

- [28] L. Applebaum, W. Bajwa, M. Duarte, and R. Calderbank. Asynchronous code-division random access using convex optimization. *Physical Communication*, 5(2):129–147, 2012.
- [29] A. Fletcher, S. Rangan, and V. Goyal. A sparsity detection framework for on-off random access channels. In *International Symposium on Information Theory*, pages 189–173, Seoul, South Korea, August 2009.
- [30] D. Guo and L. Zhang. Virtual full-duplex wireless communication via rapid on-off-division duplex. In *Allerton Conference on Communication, Control and Computing*, Monticello, IL, September/October 2010.
- [31] J. Luo and D. Guo. Neighbor discovery in wireless ad hoc networks based on group testing. In *Allerton Conference on Communication, Control and Computing*, Monticello, IL, September 2008.
- [32] J. Luo and D. Guo. Compressed neighbor discovery for wireless ad hoc networks: the Rayleigh fading case. In *Allerton Conference on Communication, Control and Computing*, Monticello, IL, September/October 2009.
- [33] L. Zhang and D. Guo. Neighbor discovery in wireless networks using compressed sensing with Reed-Muller codes. In *International Symposium on Modeling and Optimization of Mobile, Ad Hoc, and Wireless Networks*, Princeton, NJ, May 2011.
- [34] A. Thompson and R. Calderbank. Compressed neighbour discovery using sparse Kerdock matrices. In *International Symposium on Information Theory*, Vail, CO, June 2018.