Codes generated by matrix expansions

Chris Meyer

submitted in partial fulfillment of the requirements for Honors in Mathematics at the University of Mary Washington

Fredericksburg, Virginia

May 2006

This thesis by **Chris Meyer** is accepted in its present form as satisfying the thesis requirement for Honors in Mathematics.

DATE	APPROVED
______________	__________________________________
	Keith E. Mellinger, Ph.D. thesis advisor
______________	__________________________________
	Janusz Konieczny, Ph.D. committee member
______________	__________________________________
	Randall Helmstutler, Ph.D. committee member

Abstract

A new class of error-correcting codes is created from a matrix operation defined within. The matrix operation takes a point-block incidence and produces a new point-block incidence with some desirable properties, including a doubling of the girth of the initial matrix's Tanner graph. A specific example is created using $PG(2, q)$, and the results are generalized to any point-block incidence structure. These codes are analyzed mathematically and through simulation via belief propagation decoding.

1 Introduction to Error-Correcting Codes

Encoding a transmission is a method of increasing the reliability of a channel by attempting to find and correct corruptions in a signal which might occur during transmission. The applications for this technology are broad and range from compact discs and cell phones to deep space communications. In general, the idea is to add extra bits to an outgoing transmission cleverly, in a fashion that will allow the receiving station to determine the occurrence of an error, find the most likely site of the error, and possibly even correct it. This concept was introduced by Shannon around 1950 [5].

Binary linear coding is a method of implementing Shannon's ideas, and since we currently live in a world dominated by digitally represented data, the restriction to binary is logical. In this setup, encoding is accomplished via discrete packets of information of length n. A binary linear code is represented by a *generator matrix* whose entries are only 0s and 1s, a so called $(0,1)$-matrix. The message to be encoded is made up of codewords, each one a linear combination of rows from this matrix, with the addition performed modulo two. This implies that the generator matrix for a code of length n will have n columns.

The *dimension*, k, of a code is equal to the dimension of the row-space of the generator matrix. Each codeword is pre-assigned a unique meaning, so a code of dimension k is equivalent to having a 2^k messages. We often view this information in terms of the *information rate*, given by the ratio of the code's dimension to its length: $\frac{k}{n}$. The closer this ratio is to one, the more data is being passed in each packet; the closer the ratio is to zero, the more error correction information is being passed in each packet. In order to optimize the length of the codes, we would like to see codes with a high information rate which still correct many errors.

Another vital measure of a code is its *minimum distance*, d. Minimum distance measures how "far apart" the two closest codewords are, and is, in general, difficult to calculate, especially for longer length codes (in fact, this problem is known to be NP-hard). In a code with minimum distance d, the two most similar codewords will have exactly d positions different, and so the sum of these two codewords will give a codeword with exactly d 1s in it, or put otherwise, a codeword of weight d.

The code can also be represented by the *dual* of the generator matrix, also known as the *parity check matrix*. This matrix also has n columns, and every row is orthogonal to the rows of the generator matrix. Since the row-space of the parity check matrix is the dual of the code, it has dimension equal to $n - d$, by the dimension theorem from linear algebra.

For a more complete treatment of the theory of error-correcting codes, see [3], and for information on linear algebra, see [7].

2 Preliminaries

We start with some general constructions of codes using incidence structures. Here we introduce the terminology used to do so.

An *incidence structure* D is a set P of points, and a set B of blocks, where a block in B is a set of points from P. Blocks are further constrained to contain a fixed number of points, greater than or equal to two, and we assume that there are at least two blocks containing each point. Note that an incidence structure can be represented as a k-uniform *hypergraph*, which is a graph whose edges all have k vertices as constrained by the condition on blocks given above. Here we will use the terms incidence structure and hypergraph interchangeably, and mean a k-uniform hypergraph.

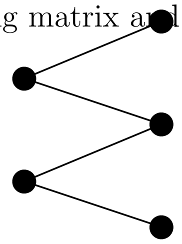
We say that a point is *incident* with a block if the block contains the point, and a *flag* is a single such incidence in D. A flag can be represented as an ordered pair (p_i, b_j), where the block b_j contains the point p_i. By assigning each point in the set P to a row of a matrix M, and each block in B to a column in M, we create an *incidence matrix* by entering 1s in the matrix entries corresponding to flags, and 0s everywhere else.

We also add some terminology borrowed from graph theory, and modify it to apply to an incidence structure or hypergraph, as defined here.

A *path* from p_1 to p_n is an alternating sequence of points and blocks p_1, b_1, p_2, b_2, ... ,b_{n-1}, p_n such that b_i contains both p_i and p_{i+1}. Just as for a regular graph, a hypergraph is *connected* if and only if for any points p_i and p_j there exists a path from p_i to p_j. With the concept of a path, we can define a *polygon*, or *n-gon*, as a path from p_i back to p_i with no blocks or points repeated except p_i. Note that an n-gon will have both n points and n blocks.

It has been conjectured that the decoding algorithm that we will use for our codes benefits from parity check matrices whose *Tanner graphs* have high *girth*, where the girth of any graph is the length of the shortest cycle in the graph, or infinity in a cycle-less graph. The Tanner graph of a matrix M, G_M, is the bipartite graph with vertex set V, one partition class corresponding to the points in D, and the other corresponding to the blocks in D. Edges between vertices in G_M exist if and only if there is a 1 in the corresponding row and column. Consider the following matrix and its Tanner graph:

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



For more information on the subject of Tanner graphs, see [6].

3 Matrix Expansion

Our general technique in constructing our codes is to take an incidence structure, apply a matrix expansion operation to it, and then to use the resulting expanded matrix as the parity-check matrix of the code.

Definition 3.1. *Given an $m \times n$ $(0,1)$-matrix M, with k non-zero entries, let the matrix $\overline{M}$ be an $(m+n) \times k$ matrix whose rows are labeled as $p_1, p_2, ...p_n, b_1, b_2, ...b_m$, and whose columns are labeled with all ordered pairs (p_i, b_j) where $M_{ij} = 1$. Furthermore, let $\overline{M}_{i,j} = 1$ if and only if the row label* *is a coordinate of the column label, and* $\overline{M}_{i,j} = 0$ *otherwise. We say that* M *has been* expanded *to* $\overline{M}$*.*

For example, below we see the expansion of a 2×3 matrix representing four flags.

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

Lemma 3.2. *If a matrix* M *represents a point-block incidence structure* D*, then a polygon of* k *sides in* D *corresponds to a cycle of length* $2k$ *in the graph* G_M*, and conversely.*

Proof. Consider a k-gon in D. By definition, there are k blocks and as noted, k points which make up this k-gon. Without loss of generality, we can describe the k-gon as a sequence of points and blocks: $\{p_1, b_1, ..., p_k, b_k\}$ where each block b_i contains the point p_i and the point p_{i+1}, with subscripts read modulo k. Note that this is a $2k$-cycle in the bipartite graph G_M.

Now consider any cycle C in the graph G_M. Since this graph is bipartite, the length of C is even. Of the $2k$ points in C, k will be points from P, and k will be blocks from B. Since edges in G_M indicate incidence, any edge in G_M corresponds to a flag in D. Thus the cycle in G_M represents a k-gon in the incidence structure. □

Theorem 3.3. *If a matrix* M *represents a point-block incidence structure, then a cycle of length* $2k$ *in* G_M *corresponds to a cycle of length* $4k$ *in the graph* $G_{\overline{M}}$*, and conversely.*

Proof. We have established that cycles in G_M have length $2k$, and represent k-gons in the incidence structure. These cycles, in general, take the form $\{p_1$, b_1, ... p_k, $b_k\}$. If we add the flags between each point and block, we have: $\{p_1$, (p_1, b_1), b_1, (p_2, b_1), p_2, ... , b_k, $(p_1, b_k)\}$. Note that this cycle of $G_{\overline{M}}$ has $4k$ unrepeated elements.

Now consider a cycle in the graph $G_{\overline{M}}$. Without loss of generality, the first point will come from the set of points, and the second from the set of flags. Then the third must come from the set of blocks, and the fourth from the set of flags again. In general, we have $\{p_1, (p_1, b_1), b_1, (p_2, b_1), p_2, ..., b_k, (p_1, b_k)\}$. If we remove the flags from this cycle, we have $\{p_1, b_1, ...p_k, b_k\}$, and also the adjacent items in this list are incident because of the flags we removed. Clearly this represents a cycle in G_M. □

Corollary 3.4. *Let k be the number of sides of the smallest polygon in an incidence structure D. Then the girth of $G_{\overline{M}}$ is $4k$.*

Proof. Let C correspond to a minimum length cycle in $G_{\overline{M}}$ with length r. By the previous theorem, C corresponds to a cycle of length $\frac{1}{2}r$ in G_M, and the corresponding cycle will have minimum length because C had minimum length. Based on the results of the lemma, the minimum length cycle of length $\frac{1}{2}r$ in G_M will correspond to a $\frac{1}{4}r$-gon in D. Letting $r = 4k$ (the smallest possible value for r), the girth of G_M is $2k$, and the girth of $G_{\overline{M}}$ is $4k$. □

While this result is neither surprising nor difficult to prove, it is of significant importance for our codes, especially considering belief propagation's conjectured preference for high girth. Consider an incidence structure where the smallest polygon is a triangle. In the expanded matrix, the girth will be twelve, a major improvement.

Since our goal is to create codes, we offer the following:

Definition 3.5. *Let M be an incidence matrix and $\overline{M}$ be its expanded matrix. We define the code $\mathcal{C}_{\overline{M}}$ to be the code generated by parity check matrix $\overline{M}$.*

4 The Code $\mathcal{C}_{\overline{\pi}}$

We now apply our results to a common incidence structure from finite geometry [2]. The code $\mathcal{C}_{\overline{\pi}}$ is derived from π, the classical projective plane of order q, also known as $PG(2,q)$. Let $PG(2,q)$ be our incidence structure with the points of the geometry as the points of D and the lines of the geometry as the blocks of D, and with incidence matrix M_π. We then expand M_π to $\overline{M_\pi}$. We will call the code with $\overline{M_\pi}$ as its parity check matrix $\mathcal{C}_{\overline{\pi}}$.

The length of the code is $q^3 + 2q^2 + 2q + 1$, following directly from the number of flags of D. Since there are $q^2 + q + 1$ points each incident with $q + 1$ lines, the number of flags is $(q + 1) \times (q^2 + q + 1) = q^3 + 2q^2 + 2q + 1$. As the flags are the columns of $\overline{M_\pi}$, the parity check matrix for $\mathcal{C}_{\overline{\pi}}$, we see that this immediately determines the length of the code.

Theorem 4.1. *The dimension of* $\mathcal{C}_{\overline{\pi}}$ *is exactly* q^3.

Proof. Recall that $\overline{M_\pi}$, the parity check matrix for $\mathcal{C}_{\overline{\pi}}$, has $2(q^2+q+1)$ rows, with q^2+q+1 of them representing points in PG(2,q), and the other q^2+q+1 representing lines in the same plane. Note that the column weight for $\overline{M_\pi}$ is exactly 2, since each column corresponds to a specific point-line flag, say (p_i, l_j), so there will be a 1 in the row corresponding to the point p_i, and another 1 in the row corresponding to the line l_j. It follows that summing the rows of $\overline{M_\pi}$ modulo 2 will give the zero vector, and hence the rows of $\overline{M_\pi}$ are linearly dependent. Now, pick an arbitrary row in $\overline{M_\pi}$. Since points and lines are interchangeable in $PG(2,q)$ (see [2]), we can assume that this row represents a point p, without loss of generality. Now we create the smallest possible linearly dependent set of rows, U, which includes this row, that is, the smallest set of rows which we can sum column-wise, with the zero vector as the result.

Since this row represents a point, it will have $q + 1$ 1s in it. In order to cancel these 1s, we must include the $q + 1$ rows corresponding to those lines, as these are the only rows which have 1s in the proper columns. By the axioms of finite projective geometry, every two lines meet in exactly one point and since these $q+1$ lines all meet in exactly one point, the q other points on each of those lines must be distinct. Hence, U must now include the $q(q+1)$ rows which correspond to these points. Notice that these q^2+q points, combined with the original point account for every point in $PG(2,q)$. Now the row-sum has a one in every column from the rows corresponding to the points, necessitating the addition of all the remaining rows corresponding to lines. Thus the smallest set of linearly dependent rows in $\overline{M_\pi}$ is in fact all of them, and therefore removing an arbitrary row will give a set of linearly independent rows. Therefore, the rank of $\overline{M_\pi}$ is $2(q^2+q+1)-1 = 2q^2+2q+1$, and so by the dimension theorem from linear algebra, the dimension of $\mathcal{C}_{\overline{\pi}}$ is $q^3+2q^2+2q+1-(2q^2+2q+1) = q^3$.

□

In order to facilitate our proofs of minimum distance (here and elsewhere), we introduce the concept of a *representative vector*. The representative vector of a collection S of points and blocks in an incidence structure D is a vector from the row space of $\overline{M}$ and has 1s in the entries corresponding the the columns in the matrix which represent the flags included in the collection S. This definition is admittedly awkward, and so we include an example for understanding.

Consider $M = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{bmatrix}$, and $\overline{M} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$. The representative vector of the set p_1, b_1, p_2 would consist of the flags (p_1, b_1), (p_2, b_1) and be equal to [1100].

Lemma 4.2. *The representative vector of an n-gon in the incidence structure D is a codeword of weight $2n$ in $\mathcal{C}_{\overline{M}}$.*

Proof. Let V be the representative vector of an n-gon from D. In an n-gon, each point p_i is incident with exactly two blocks, so there are exactly two flags in V which have p_i as a coordinate. Since there are n points in the n-gon, there are exactly $2n$ flags in the representative vector. Notice that each block b_j is also incident with exactly two points, so similarly there will be exactly two flags in V which have b_j as a coordinate. Consider an arbitrary row U in $\overline{M}$. Assuming U represents a point p_i, if p_i is not in the n-gon, none of the flags will have p_i as a coordinate, and hence U will have zero 1s in common with V. If, on the other hand, p_i is in the n-gon, then as noted there will be two flags in V with p_i in their coordinates, but U will also have those same two 1s, so U and V share an even number of 1s. Similarly, if U were to represent a block, U would share an even number of 1s with V. Therefore V is orthogonal to every row of $\overline{M}$, and is a codeword for $\mathcal{C}_{\overline{M}}$.

□

Theorem 4.3. *The minimum distance of* $\mathcal{C}_{\overline{\pi}}$ *is* 6.

Proof. We show the upper bound on minimum distance by exhibiting a codeword of weight 6. In $PG(2, q)$, it is known that triangles exist. In a triangle, there are 3 points, each incident with 2 lines. Hence a triangle has 6 flags. By the previous lemma, every triangle represents a codeword of weight 6.

We show the lower bound by constructing the smallest possible non-zero codeword. Let c be the smallest possible codeword. Since c is non-zero, without loss of generality, we can say that c has a one in the column corresponding to (p_1, l_1). Since c is a codeword, c is orthogonal to every row of M_π, specifically the row corresponding to p_1. Since c shares a single 1 with this row in the column (p_1, l_1), then c must also share another 1 with this row, and without loss of generality, that 1 is in the column (p_1, l_2). Based on this information, we know that c shares a 1 with both the rows l_1 and l_2, and so must also share a second one with each of these rows. It is impossible that these two lines could be incident with another point, because every two lines determine exactly one point, and l_1 and l_2 determine p_1. So l_1 is incident with some point $p_2 \neq p_1$, and so both of those rows have a 1 in the column corresponding to (p_2, l_1). Now c is orthogonal to the row l_1, though the row p_2 must share some other 1 with c. Furthermore, it is impossible that p_2

could be incident with l_1, because we would arrive at the same contradiction as before. So p_2 must be incident with l_3, and now, having a 1 in the column (p_2, l_3), the row p_2 is orthogonal to c. Both of the rows l_2 and l_3 need another 1 in order to be orthogonal to c. Let p_3 be incident with both these lines. Now c has a 1 in the columns corresponding to (p_1, l_1), (p_1, l_2), (p_2, l_1), (p_2, l_3), (p_3, l_2), (p_3, l_3), and a zero in every other column, and thus has weight 6. □

As outlined in this section, $\mathcal{C}_{\overline{\pi}}$ is a $[q^3 + 2q^2 + 2q + 1, q^3, 6]$ code. The information rate, $\frac{n}{k}$, is very high for $\mathcal{C}_{\overline{\pi}}$, in fact, it approaches 1 as q grows. However, the minimum distance is fixed at 6, the price for such a high information rate.

5 General Results on $\mathcal{C}_{\overline{M}}$

As mentioned before, the results we have obtained for $\mathcal{C}_{\overline{\pi}}$ can be applied to incidence structures in general.

For an arbitrary incidence structure D with incidence matrix M, the length of $\mathcal{C}_{\overline{M}}$ will always be exactly the number of flags of the incidence structure, the number of 1s in the matrix M or the number of columns in the matrix $\overline{M}$. Let B be the set of blocks of D, P be the set of points of D, and F the set of flags of D.

Theorem 5.1. *If D is connected, then the dimension of $\mathcal{C}_{\overline{M}}$ is exactly $k = |F| - |B| - |P| + 1$*

Proof. We create the smallest possible linearly dependent set of rows of $\overline{M}$. Assume, without loss of generality that p_1 is in this set. As D is connected, then for any i, there exists a path from p_1 to p_i. Since p_1 is in this set, the only way to cancel out all of the 1's in the row corresponding to p_1 is to introduce all of the rows corresponding to the blocks which contain p_1, including b_1. Now, to cancel out the 1s in the row corresponding to b_1, we must introduce all of the rows corresponding to the points contained in b_1, including p_2. In this manner, the rows corresponding to the members of the path from p_1 to p_i must be included in our linearly dependent set of rows. Since p_i is arbitrary we must include every point, and because every block contains at least one point, then we must include all of the rows corresponding to blocks. Then the smallest possible linearly dependent set of rows, is in fact, all of them. So the largest set of independent rows of $\overline{M}$ is all of the rows, minus any one. So the rank of $\overline{M}$ is $|B| + |P| - 1$. Thus the dimension of $\mathcal{C}_{\overline{M}}$ is $|F| - [|B| + |P| - 1]$.

□

Porism 5.2. *If D has K components, then the dimension of $\mathcal{C}_{\overline{M}}$ is exactly $|F| - [|B| + |P| - K]$*

Proof. Let $C_1, C_2, ..., C_K$ be the K connected components of D. As seen in the proof of the last theorem, the rows of a connected component are independent if one row is removed. Since none of the components can possibly interact to create a linearly dependent set of rows, removing one row from each component (K rows in total) will leave behind a maximal set of independent rows. □

Theorem 5.3. *The minimum distance of $\mathcal{C}_{\overline{M}}$ is exactly $2k$, where k is the size of smallest polygon in D.*

Proof. We create the smallest codeword in $\mathcal{C}_{\overline{M}}$, say c. Without loss of generality, c contains the flag (p_1, b_1). So now c has a single 1 in common with both the rows p_1 and b_1. The flag (p_1, b_2) must be added to c so that row p_1 shares an even number of 1's with c. Likewise, we must add (p_2, b_1) so row b_1 will be orthogonal to c. Thus any point in D which shares a flag with c must in fact share two flags, hence there are two blocks included for each point. Similarly, any block which shares a flag with c must actually share two flags, so there will be two points included for each block included. Clearly then, c must contain a set of flags which forms a polygon: this is the only way to guarantee that each block contains two points from c, and each point is contained by two blocks from c. Since c is the least-weight codeword, c must contain the flags corresponding to the smallest polygon, say $(p_1, b_1), (p_2, b_1), (p_2, b_2), ..., (p_k, b_k), (p_1, b_k)$. This set clearly has $2k$ elements. Thus the minimum distance of $\mathcal{C}_{\overline{M}}$ is $2k$ where k is the number of sides in the smallest polygon in D. □

With $\mathcal{C}_{\overline{M}}$, codes with virtually any parameters can be generated by choosing M in a clever fashion. Perhaps the most notable feature follows from Corollary 3.4, which explains how the girth of a Tanner graph doubles after matrix expansion. For decoding algorithms which favor high girth, this provides an easy way to capitalize on that advantage.

6 Simulation Data

To demonstrate the effectiveness of the codes we have developed, we used an iterative probabilistic decoding algorithm published in [4] and freely available on the Internet[1]. The algorithm "sends" a large number of randomly generated codewords with errors, then attempts to decode them using belief propagation and counts the number of errors that are still present after a certain number of iterations. The initial incidence structures were created with the software package Magma [1].

As an aid to understanding, Figure 1 shows the performance of five codes of various length, all generated by the method described in Section 3, the code $\mathcal{C}_{\overline{\pi}}$. Along the x-axis, we have the signal to noise ratio, or simply the relative signal strength, where 6 is a relatively strong signal, and 1 is relatively noisy. On the y-axis we have the rate of errors getting past the error correction at any given signal to noise ratio. Note that the y-axis is a logarithmic scale, indicating that dropping one major unit on the axis is equivalent to a ten-times decrease in errors. For comparison, the dashed line indicates an uncoded signal run through the decoding algorithm. We conclude that our

[1] http://the-art-of-ecc.com/8_Iterative/BPdec/pearl.c

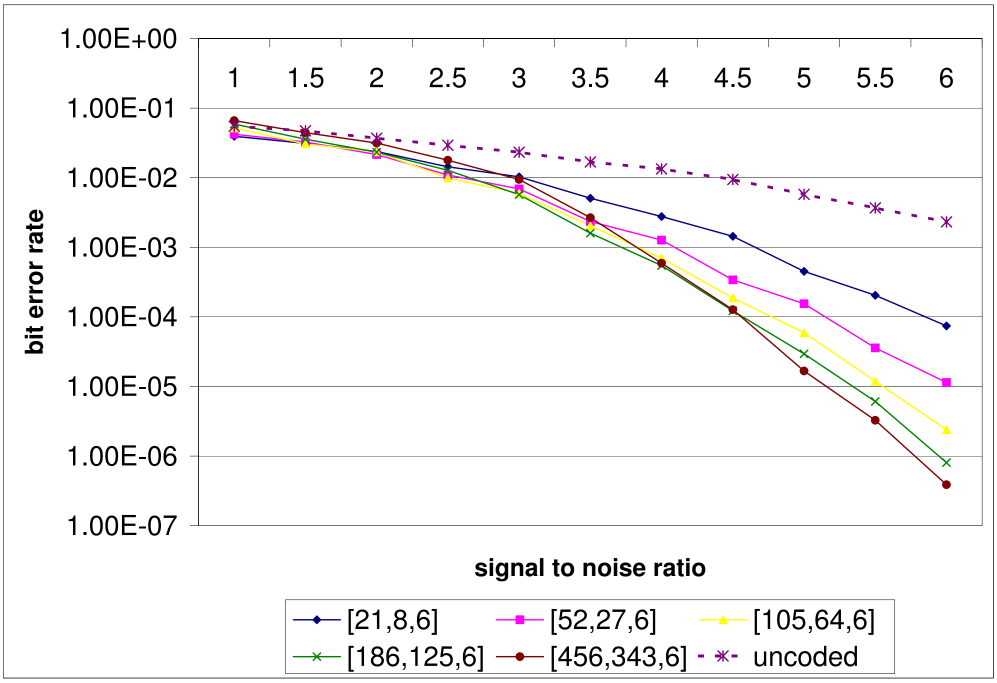


Figure 1: Performance chart for codes $\mathcal{C}_\pi$

codes are performing at an acceptable level, as they consistently outperform the uncoded signal.

References

[1] J. Cannon and C. Playoust, “An Introduction to Magma”, University of Sydney, Sydney, Australia (1994).

[2] J.W.P. Hirschfeld, “Projective Geometries over Finite Fields,” Oxford University Press, second edition (1998).

[3] W. C. Huffman and V. Pless, “Fundamentals of Error-Correcting Codes,” Cambridge University Press (2003).

[4] R.H. Morelos-Zaragoza, The Art of Error Correcting Coding, Wiley, 2002.

[5] C. E. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.* **27** (1948), 379–423, 623–656.

[6] R. M. Tanner, A recursive approach to low complexity codes, *IEEE Trans. Inform. Theory* **IT - 27** (1981), 533–547.

[7] S. Venit and W. Bishop, “Elementary Linear Algebra,” Brooks/Cole Publishing Company, fourth edition (1995).