# Securing Data Transmission Using Coding Theory

Feryâl Alayont
*Department of Mathematics*
*University of Arizona*

April 11, 2006

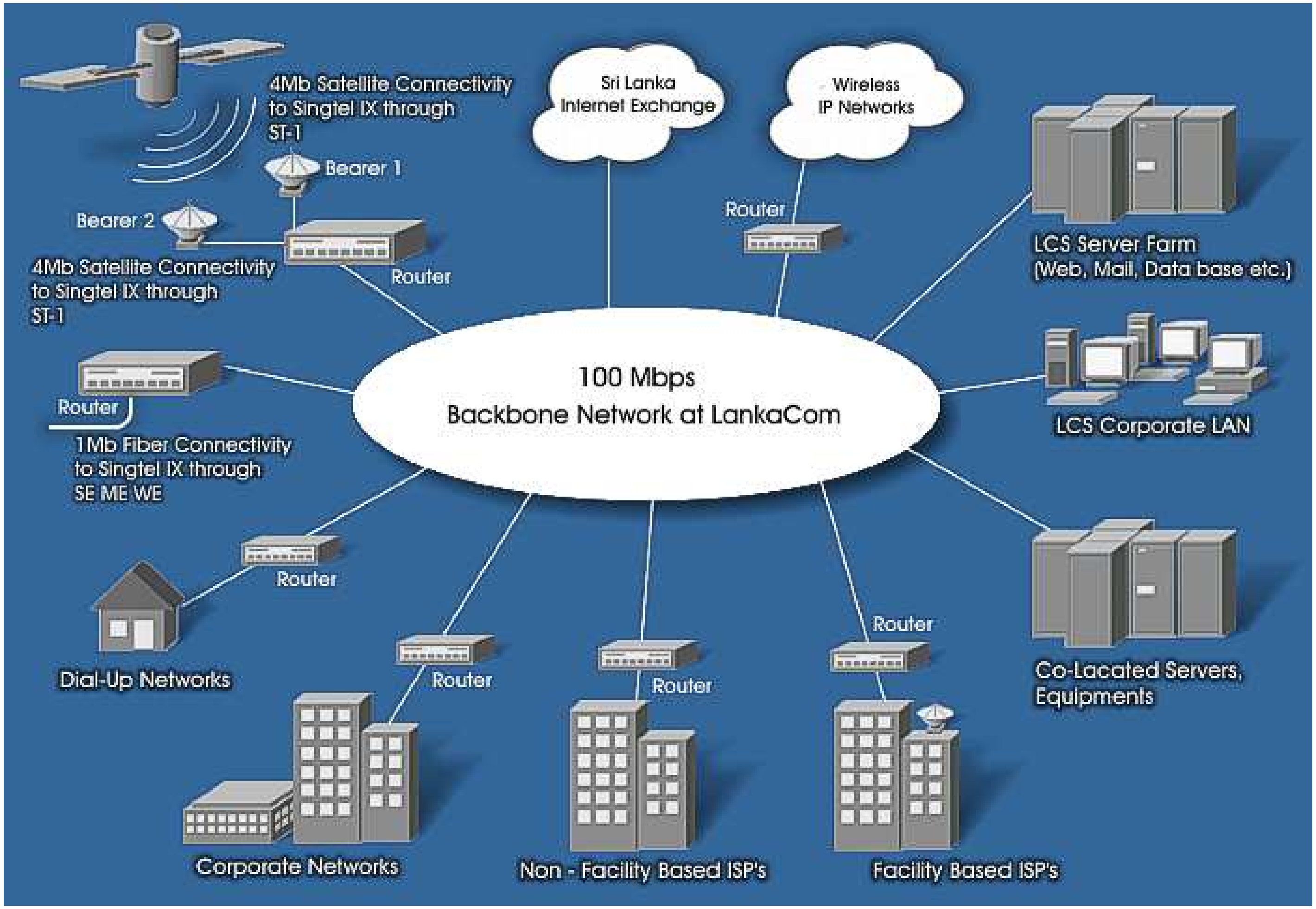



Figure 1: A network connectivity example, http://www.lankacom.net/images/internetwork.htm

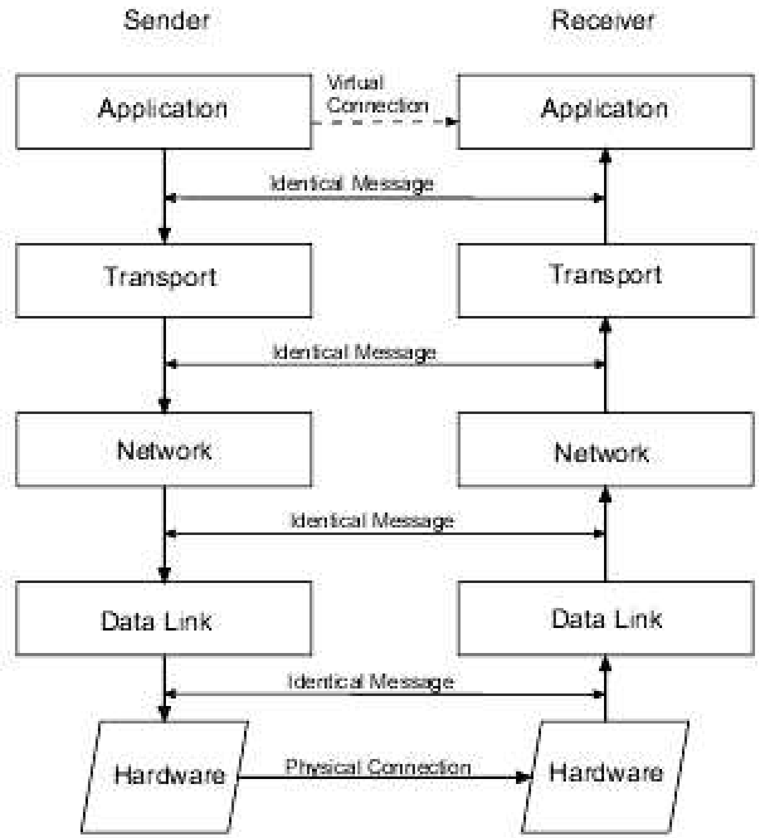
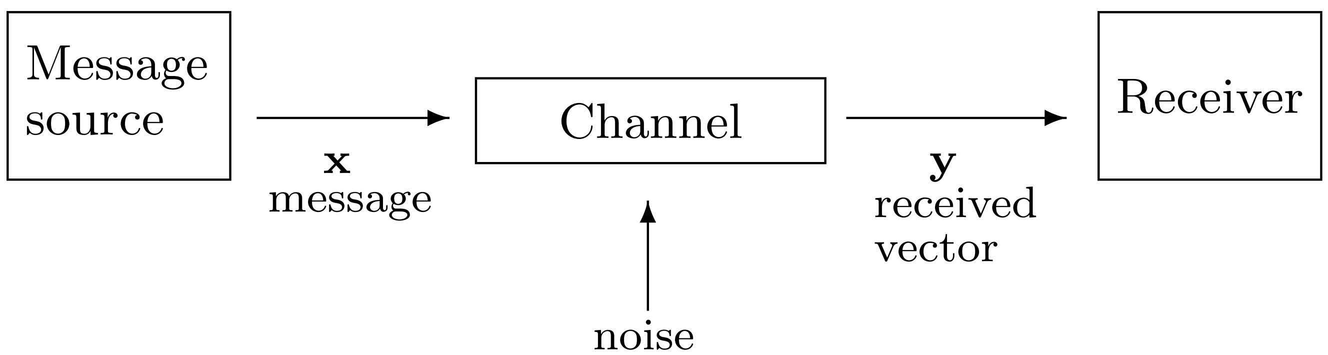



Figure 2: Flow of data between two computers, http://www.rabbitsemiconductor.com/documentation/docs/manuals/TCPIP/Introduction/4layers.htm

Data transmission examples that use coding theory: Wireless communication, CD burning/reading, satellite communication, space missions, ...

Communication channel:



Messages in binary digits:

$$\text{Sent: } 0111 \xrightarrow{\text{noisy channel}} \text{Received: } 0101$$

Error not even detected!

Solution?

**Error detection:** Repeat messages twice

$$\text{Message: } \mathbf{x} = 0111 \quad \rightsquigarrow \quad \text{Sent: } \mathbf{c} = 0111|0111$$
$$\xrightarrow{\text{noisy channel}} \text{Received: } \mathbf{y} = 0101|0111$$

The two parts don't match! (Single) error detected!

$$\textit{Information rate} = \frac{\text{length of message}}{\text{length of sent word}} = \frac{1}{2}$$

**Better detection method:** Overall parity check (checksum)

Append a digit to the end so that total number of 1's is even

Mathematically: $\mathbf{x} = x_1x_2x_3x_4$ is coded as $\mathbf{c} = x_1x_2x_3x_4x_5$ so that $x_1 + x_2 + x_3 + x_4 + x_5 = 0 \mod 2$

$$\text{Message: } \mathbf{x} = 0111 \qquad 0+1+1+1+x_5 = 0 \mod 2$$

$$\text{Sent: } \mathbf{c} = 0111|1 \xrightarrow{\text{noisy channel}} \text{Received: } \mathbf{y} = 0101|1$$

Parity check: $0+1+0+1+1 \neq 0 \mod 2$

Parity check doesn't work! (Single) error detected!

$$\textit{Information rate} = \frac{\text{length of message}}{\text{length of sent word}} = \frac{4}{5}$$

**Single error correction:** Repeat messages three times

$$\text{Message: } \mathbf{x} = 0111 \quad \rightsquigarrow \quad \text{Sent: } \mathbf{c} = 0111|0111|0111$$
$$\xrightarrow{\text{noisy channel}} \text{Received: } \mathbf{y} = 0111|0101|0111$$

Choose the part which is repeated at least two times. Single error corrected!

$$\textit{Information rate} = \frac{\text{length of message}}{\text{length of sent word}} = \frac{1}{3}$$

**Better error-correcting code:** Hamming [7,4] code; a single-error-correcting code

Add 3 bits $x_5, x_6, x_7$ to the message $x_1 x_2 x_3 x_4$ so that

$$\begin{aligned} x_2 + x_3 + x_4 + x_5 &= 0 \mod 2 \\ x_1 + x_3 + x_4 + x_6 &= 0 \mod 2 \\ x_1 + x_2 + x_4 + x_7 &= 0 \mod 2 \end{aligned}$$

Matrix notation:

$$\begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

The Hamming [7,4] code is the kernel (null space) of this matrix.

Message: $\mathbf{x} = 0111$

$1+1+1+x_5 = 0,\ 0+1+1+x_6 = 0,\ 0+1+1+x_7 = 0$

Sent: $\mathbf{c} = 0111|100 \xrightarrow{\text{noisy channel}}$ Received: $\mathbf{y} = 0101|100$

Which parity check equations are not satisfied?

Recall:

$$
\begin{aligned}
x_2 + x_3 + x_4 + x_5 &= 0 \mod 2 \\
x_1 + x_3 + x_4 + x_6 &= 0 \mod 2 \\
x_1 + x_2 + x_4 + x_7 &= 0 \mod 2
\end{aligned}
$$

$1+0+1+1 \neq 0 \qquad 0+0+1+0 \neq 0 \qquad 0+1+1+0 = 0$

Message: $\mathbf{x} = 0111$

$1 + 1 + 1 + x_5 = 0,\ 0 + 1 + 1 + x_6 = 0,\ 0 + 1 + 1 + x_7 = 0$

Sent: $\mathbf{c} = 0111|100 \xrightarrow{\text{noisy channel}}$ Received: $\mathbf{y} = 0101|100$

Which parity check equations are not satisfied?

Recall:

$$\begin{aligned} x_2 + x_3 + x_4 + x_5 &= 0 \mod 2 \\ x_1 + x_3 + x_4 + x_6 &= 0 \mod 2 \\ x_1 + x_2 + x_4 + x_7 &= 0 \mod 2 \end{aligned}$$

$$1 + 0 + 1 + 1 \neq 0 \qquad 0 + 0 + 1 + 0 \neq 0 \qquad 0 + 1 + 1 + 0 = 0$$

3rd position is where the error is! Correct: $\hat{\mathbf{c}} = 0111|100$

Another example: Received: 1001|110

Which parity check equations are not satisfied?
Recall:

$$
\begin{aligned}
x_2 + x_3 + x_4 + x_5 &= 0 \mod 2 \\
x_1 + x_3 + x_4 + x_6 &= 0 \mod 2 \\
x_1 + x_2 + x_4 + x_7 &= 0 \mod 2
\end{aligned}
$$

$$0 + 0 + 1 + 1 = 0 \qquad 1 + 0 + 1 + 1 \neq 0 \qquad 1 + 0 + 1 + 0 = 0$$

Another example: Received: 1001|110

Which parity check equations are not satisfied?
Recall:

$$\begin{aligned} x_2 + x_3 + x_4 + x_5 &= 0 \mod 2 \\ x_1 + x_3 + x_4 + x_6 &= 0 \mod 2 \\ x_1 + x_2 + x_4 + x_7 &= 0 \mod 2 \end{aligned}$$

$$0+0+1+1=0 \qquad 1+0+1+1 \neq 0 \qquad 1+0+1+0=0$$

6th position is where the error is! Correct: $\hat{\mathbf{c}} = 1001|100$

$$\textit{Information rate for the Hamming code} = \frac{\text{length of message}}{\text{length of sent word}} = \frac{4}{7}$$

*Linear code,* $\mathcal{C}$: set of binary codewords which includes the codeword with all 0's and coordinatewise sum of any two codewords

Example: 0000    1000    0111    1111

Example: Codewords satisfying $Hx = 0$
$H(0, 0, \ldots, 0) = 0$
$Hx_1 = 0$ and $Hx_2 = 0 \Longrightarrow H(x_1 + x_2) = 0$

**Note:** A linear code is a subspace in $F_2^n$

Basis of $\mathcal{C}$: $r_1$, $r_2, \ldots r_k$, $k$=dimension of $\mathcal{C}$

Generator matrix of $\mathcal{C}$: $G = \begin{bmatrix} -r_1- \\ -r_2- \\ \cdots \\ -r_k- \end{bmatrix}$

Encoding: $\mathbf{x} \leadsto \mathbf{c} = \mathbf{x}G$

Example: 000000 100011 010101 001110

110110 101101 011011 111000

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

$$\begin{array}{ll} 000 \leadsto 000000 & 100 \leadsto 100011 \\ 010 \leadsto 010101 & 001 \leadsto 001110 \\ 110 \leadsto 110110 & 011 \leadsto 011011 \\ 101 \leadsto 101101 & 111 \leadsto 111000 \end{array}$$

Hamming distance: $d(\mathbf{x}, \mathbf{y})$ =number of coordinates in which $\mathbf{x}$ and $\mathbf{y}$ differ

Weight: $d(\mathbf{x}, 0) = wt(\mathbf{x})$

Examples: $d(0000, 0011) = 2, \quad d(0000, 1010) = 2,$
$d(0000, 1011) = 3$

Hamming distance is a distance function, in particular the Triangle Inequality holds:

$$d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$$

Decoding: $\mathbf{y} = \mathbf{c} + \mathbf{e}$, $\mathbf{e}$ = error vector

Strategy: guess that the codeword sent is the codeword $\hat{\mathbf{c}}$ such that the number of errors is minimum

$\iff$ $\mathbf{e} = \mathbf{y} - \hat{\mathbf{c}}$ has the least number of 1's

$\iff$ $wt(\mathbf{e}) = d(\mathbf{y}, \hat{\mathbf{c}})$ is minimum

$\iff$ $\hat{\mathbf{c}}$ is the nearest codeword neighbor of $\mathbf{y}$

Sent: $\mathbf{c} \rightsquigarrow$ Received: $\mathbf{y}$

Decode: $\hat{\mathbf{c}}$ = nearest codeword neighbor of $\mathbf{y}$

Example: Code

000000 100011 010101 001110

110110 101101 011011 111000

Received: 011010 $\rightsquigarrow$ Nearest neighbor: 011011

Decode: 011011

Received: 101010 $\rightsquigarrow$ Nearest neighbor: 100011 or 001110 or 111000 ?

Cannot be determined

Which errors can be corrected?

Minimum distance of a code $\mathcal{C}$ is the minimum distance between two distinct words in the code.

Example: 000000    100011    010101    001110
110110    101101    011011    111000
has minimum distance 3: $d(000000, 100011) = 3$.

**Theorem.** *If $\mathcal{C}$ is a code with minimum distance $d$, nearest neighbor decoding correctly decodes any received vector in which at most $\lfloor \frac{d-1}{2} \rfloor$ errors have occurred.*

***Proof:*** Sent: $\mathbf{c}$ Error: $\mathbf{e}$ with less than $\lfloor \frac{d-1}{2} \rfloor$ 1's
Received: $\mathbf{y} = \mathbf{c} + \mathbf{e}$

Claim: $\mathbf{c}$ is the unique codeword closest to $\mathbf{y}$.

If $\mathbf{c}'$ is another codeword with distance at most $\lfloor \frac{d-1}{2} \rfloor$ from $\mathbf{y}$:

$$d(\mathbf{c}, \mathbf{c}') \leq d(\mathbf{c}, \mathbf{y}) + d(\mathbf{y}, \mathbf{c}') \leq \lfloor \frac{d-1}{2} \rfloor + \lfloor \frac{d-1}{2} \rfloor \leq d - 1$$

Contradiction. $\square$

Which errors cannot be corrected?

Received $\mathbf{y} = \mathbf{c} + \mathbf{e} = \mathbf{c}' + \mathbf{e}'$ for $\mathbf{c} \neq \mathbf{c}'$. Decoded as $\mathbf{c}$.

$\mathbf{e}' = \mathbf{e} + (\mathbf{c} - \mathbf{c}') \in \mathbf{e} + \mathcal{C}$ and $wt(\mathbf{e}) < wt(\mathbf{e}')$.

All possible errors$=F_2^n$ decomposes into cosets of $\mathcal{C}$. All errors in a coset are decoded as the minimal weight error in that coset.

Example:

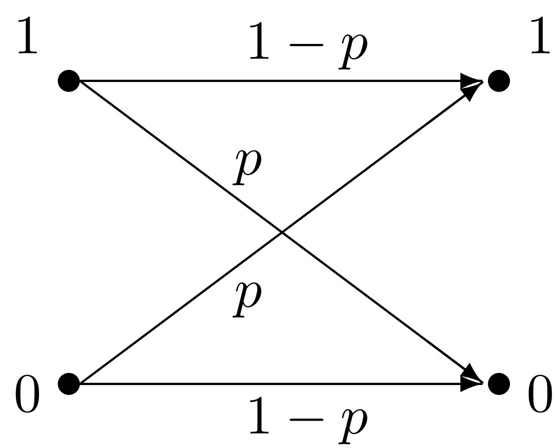
| | Leader | |
|---|---|---|
| Code | 000 | 111 |
| | 100 | 011 |
| | 010 | 101 |
| | 001 | 110 |

Coset leaders for Hamming $[7, 4, 3]$ code:

| | Leader | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Code | 0000000 | 0010110 | 0100101 | 0110011 | 1000011 | 1010101 | 1100110 | 1110000 |
| | 1000000 | 1010110 | 1100101 | 1110011 | 0000011 | 0010101 | 0100110 | 0110000 |
| | 0100000 | 0110110 | 0000101 | 0010011 | 1100011 | 1110101 | 1000110 | 1010000 |
| | 0010000 | 0000110 | 0110101 | 0100011 | 1010011 | 1000101 | 1110110 | 1100000 |
| | 0001000 | 0011110 | 0101101 | 0111011 | 1001011 | 1011101 | 1101110 | 1111000 |
| | 0000100 | 0010010 | 0100001 | 0110111 | 1000111 | 1010001 | 1100010 | 1110100 |
| | 0000010 | 0010100 | 0100111 | 0110001 | 1000001 | 1010111 | 1100100 | 1110010 |
| | 0000001 | 0010111 | 0100100 | 0110010 | 1000010 | 1010100 | 1100111 | 1110001 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 0001111 | 0011001 | 0101010 | 0111100 | 1001100 | 1011010 | 1101001 | 1111111 |
| 1001111 | 1011001 | 1101010 | 1111100 | 0001100 | 0011010 | 0101001 | 0111111 |
| 0101111 | 0111001 | 0001010 | 0011100 | 1101100 | 1111010 | 1001001 | 1011111 |
| 0011111 | 0001001 | 0111010 | 0101100 | 1011100 | 1001010 | 1111001 | 1101111 |
| 0000111 | 0010001 | 0100010 | 0110100 | 1000100 | 1010010 | 1100001 | 1110111 |
| 0001011 | 0011101 | 0101110 | 0111000 | 1001000 | 1011110 | 1101101 | 1111011 |
| 0001101 | 0011011 | 0101000 | 0111110 | 1001110 | 1011000 | 1101011 | 1111101 |
| 0001110 | 0011000 | 0101011 | 0111101 | 1001101 | 1011011 | 1101000 | 1111110 |

Binary symmetric channel:




$p$: probability of bit error

Probability of word error, $P_{err}$ = probability of incorrect or ambiguous decoding

$\iff$ $P_{err}$ = probability of the error not being a coset leader

Probability of a particular error of weight $i = p^i(1-p)^{n-i}$ because $i$ errors occurred

Probability of the error being a coset leader $= \sum_i$ probability of the error being a coset leader of weight $i$
$= \sum_i \alpha_i p^i (1-p)^{n-i}$ where $\alpha_i$ is the number of coset leaders of weight $i$

$$P_{err} = 1 - \sum_i \alpha_i p^i (1-p)^{n-i}$$

Example: $P_{err} = 1 - (1-p)^4$ for sending length $n = 4$ words without encoding

$P_{err} = 1 - (1-p)^7 - 7p(1-p)^6$ for Hamming $[7, 4, 3]$ code

For $p = 1/100$, the first is $\approx 0.0394$ and the second is $\approx 0.0020$.

For a binary symmetric channel with probability of bit error $0 < p < 1$, the channel capacity is

$$C = 1 + p\log_2(p) + (1-p)\log_2(1-p)$$

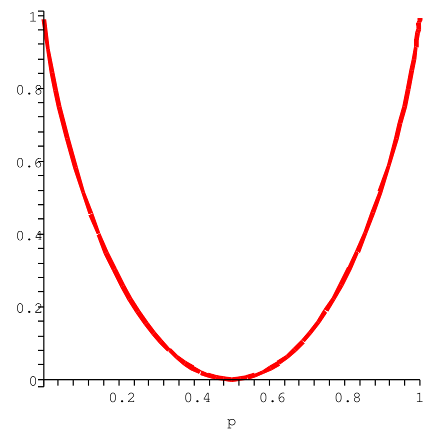



Figure 3: Channel capacity function

**Theorem.** *(Shannon, 1948) Given* $\epsilon > 0$ *and* $R < C$*, there exists a sufficiently long linear code with rate greater than* $R$ *and probability of decoding error less than* $\epsilon$*. No such linear code exists if* $R > C$*.*

Good codes: RSV codes, Low-density parity-check codes, turbo codes

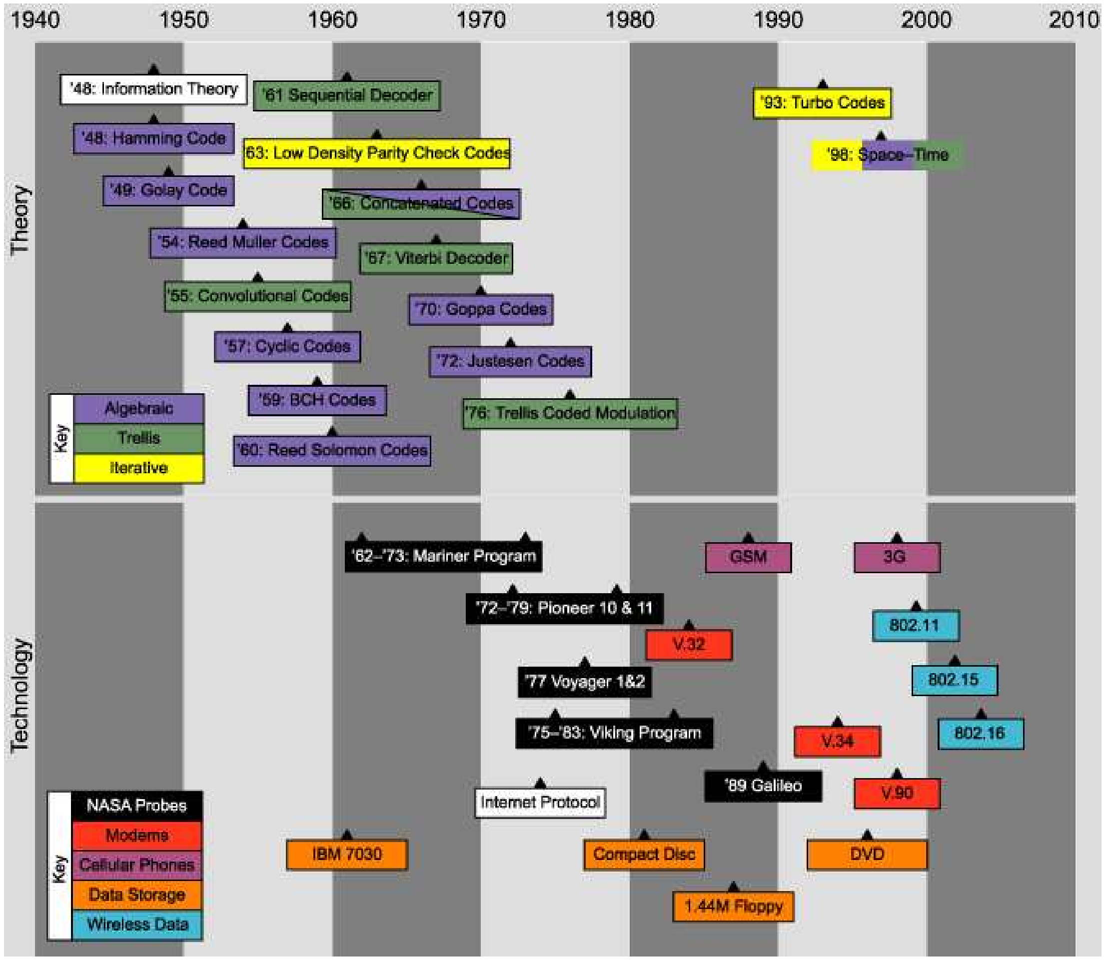



Figure 4: Timeline of error control coding, http://www.acorn.net.au/telecoms/coding/coding.cfm

## References:

Fundamentals of Error-Correcting Codes, W. Cary Huffman, Vera Pless, Cambridge University Press, 2003.

Introduction to Coding Theory, J.H. Van Lint, Springer, 1998.

Introduction to the Theory of Error-Correcting Codes, Vera Pless, Wiley-Interscience, 1989.

The Mathematics of Coding Theory, Paul Garrett, Prentice Hall, 2003.

The Theory of Error-Correcting Codes, F.J. MacWilliams, N.J.A. Sloane, North-Holland Mathematical Library, 1977.

ARC Communications Research Network page on Error Control Coding,
http://www.acorn.net.au/telecoms/coding/coding.cfm

Neil J.A. Sloane's webpage
http://www.research.att.com/~njas

Wikipedia page on Internet,
http://en.wikipedia.org/wiki/Internet

Wikipedia page on Internet Protocol,
http://en.wikipedia.org/wiki/Internet_Protocol