Compression, Correction, Confidentiality, and Comprehension

Steven M. Bellovin
smb@cs.columbia.edu
http://www.cs.columbia.edu/~smb

+1 212-939-7149
Department of Computer Science
Columbia University
Early Telegraphy

• Early telegraphy, especially overseas, was very expensive: $100 for twenty words trans-Atlantic in 1866.

• Messages were no longer sealed; a telegraph operator saw them

• The solution was code books

• Precedents: optical semaphore networks; naval signaling flags
Questions

• What did code books *really* do?

• How did they fit into the overall communications picture?

• Is there a relationship to modern network technology?
The Network Stack

- Standardized model; doesn’t completely match the Internet
- Each layer provides services to the layer above
- Layers rely on the properties of the layer below
- Layers communicate with their peers on other nodes
The *Real* Network Stack

<table>
<thead>
<tr>
<th></th>
<th>Political</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>Financial</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Application</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Presentation</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Session</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Transport</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Network</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Link</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Physical</td>
<td></td>
</tr>
</tbody>
</table>
The Telegraph Stack

Note: link/network layers merged here; there could be many transmissions over different telegraph links
Decoding During Transmission

Some companies offered in-net decoding, to avoid problems from code book mismatches.
Fitting Four Focus Areas to the Stack

- Compression — reducing transmission cost
- Correction — detecting and correcting errors
- Confidentiality — protecting the content of a message
- Comprehension — understanding other cultures, distant in time and space
Compression
Compression Metrics

• The goal was not to minimize characters sent, it was to minimize cost

  ➡ A layer 8 consideration

• Cost was affected by telegraph company tariffs and international regulations

  ➡ Layer 9?

• Permissible “words” changed over time: words in the local language, words in one of several languages, pseudo-words that were “pronounceable”, ten letters with a certain vowel density — and ultimately, any five-letter sequence
Domain-Specific Compression

• Many professions had their own code books

• Even explosives manufacturers had their own code

• Example: in the *The Theatrical Cipher Code* (1905), DISORB meant do not want drunkards and FILIATION meant chorus girls who are shapely and good looking

• We still use domain-specific compression: Lempel-Ziv does not work nearly as well as JPEG and MP3 on pictures or audio files
The Theatrical Cipher Code (1905)

Filacer.......An opera company
Filament.......Are they willing to appear in tights
Filander.......Are you willing to appear in tights
Filar.........Ballet girls
Filaria.......Burlesque opera
Filature.......Burlesque opera company
File.........Burlesque people
Filefish.......Chorus girl
Fileial........Chorus girls
Filially.......Chorus girls who are
Filiation.......Chorus girls who are shapely and good looking
Filibuster.....Chorus girls who are shapely, good looking and can sing
Filicoid.......Chorus girls who can sing
Filiform.......Chorus man
Filigree.......Chorus men
Filing........Chorus men who can sing
Fillet.........Chorus people
Fillip.........Chorus people who can sing
Filly.........Comic opera
Film..........Comic Opera Company
Filter........Comic Opera people
Filtering.......Desirable chorus girl
Unofficial Navy Code (1909)

For the 2008 game, you would send

**FEPUB** BAWAS: Army 0, Navy 34
Correction
Error Correction

• What about errors during transmission? Errors are link layer-specific, and a given message could be sent over multiple link types.

• In the police code, SUB is Vienna, but SYB is Jerusalem — and U and Y are adjacent on the keyboard.

• Morse code had its own errors. Consider how .._. (F) could be received:

  • IN (.. __.)
  • ER (._._.)
  • UE (..___.)

Wednesday, October 14, 2009
Techniques

• Terminal indices

• Check digits

• Two-letter differences

• Avoidance of common words

• Mutilation Tables
Mutilation Tables

Look up the first two letters in the upper left, move across to the middle letter, move down to the lower table. Context often permits disambiguation of the possible original words from the various legal possibilities.
Confidentiality
Add or subtract a prearranged key; monoalphabetic substitution of letter.

**The Secret Corresponding Vocabulary (1845)**

<table>
<thead>
<tr>
<th>1551 Chorus</th>
<th>1601 Chronological</th>
<th>1651 Churlish</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 Chose</td>
<td>2 ally</td>
<td>2 ishly</td>
</tr>
<tr>
<td>3 en</td>
<td>3 Chronometer</td>
<td>3 ishness</td>
</tr>
<tr>
<td>4 Chouse</td>
<td>4 ric</td>
<td>4 ly</td>
</tr>
<tr>
<td>5 ed</td>
<td>5 rical</td>
<td>5 Churn</td>
</tr>
<tr>
<td>6 ing</td>
<td>6 etry</td>
<td>6 ed</td>
</tr>
<tr>
<td>7 Chowder</td>
<td>7 Chrysalis</td>
<td>7 ing</td>
</tr>
<tr>
<td>8 Christ</td>
<td>8 Chrysography</td>
<td>8 staff</td>
</tr>
<tr>
<td>9 less</td>
<td>9 Chrysolite</td>
<td>9 Chyle</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1560 Christen</th>
<th>1610 Chub</th>
<th>1660 ifaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 dom</td>
<td>1 bed</td>
<td>1 ifactive</td>
</tr>
<tr>
<td>2 ed</td>
<td>2 by</td>
<td>2 iferous</td>
</tr>
<tr>
<td>3 ing</td>
<td>3 faced</td>
<td>3 ous</td>
</tr>
</tbody>
</table>
Threat Models

On the 1st February, 1870, the telegraph system throughout the United Kingdom passes into the hands of the Government, who will work the lines by Post Office officials. In other words, those who have hitherto so judiciously and satisfactorily managed the delivery of our sealed letters will in future be entrusted also with the transmission and delivery of our open letters in the shape of telegraphic communications, which will thus be exposed not only to the gaze of public officials, but from the necessity of the case must be read by them.

Slater's Telegraphic Code (1870)
Slater’s Telegraph Code (1870-1939)

- Long-lived

- Encode to 5-digit numbers

- Use additives, transpositions of digits, or combinations

- Map result to other code words

- Note: the resulting message was quite expensive: there was no error detection or compression, and the code words were expensive under later tariffs. But the code lasted for almost 70 years.
Bloomer's Commercial Cryptograph: A Telegraph Code and Double Index—Holocryptic Cipher (1874)

- Holocryptic: “wholly hidden or secret; spec. of a cipher incapable of being read except by those who have the key” (OED)

- Standard code words, code numbers, and phrases

- Suggestions for additives, transposition of code words, and user-generated two-part code variant

- Different additives could be used for different words (the “holocryptic” part)

- Room for user-created two-part codes (“double index”)
INSTRUCTIONS.

This Cipher Code arranged for use of the several Organizations of Railway Employes is intended more especially for Telegraphic Correspondence in time of trouble, when it is desirable or necessary to send telegrams that can not be read by any but those for whom they are intended, as is the case in time of strikes or other important moves on the part of an Organization, as it is often necessary to use the Company’s wire to reach members of the Organization on other parts of the line.

Labor versus Management

Labor had more secure codes...
INSTRUCTIONS

This Code will be designated by the word VAN, and is to be used only when secrecy is desired.

If the entire message is in cipher, the word VAN must begin and end the message.

It may frequently be deemed unnecessary to cipher every word. When only part of a message is ciphered, the ciphered word or words must be preceded and followed by the word VAN.

The NY Central’s VAN Code (1923)
Governments Were No Better

“When a single key number is used, the number may be alternately added and subtracted. Other methods will readily occur. The use of 50 or 100, while easy to remember, should be avoided.”

U.S. War Department, 1904
## Wiring Money: A Two-Part Code (1952)

### Identification

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chain</td>
<td>190</td>
<td>Gable</td>
<td>128</td>
<td>Maxim</td>
<td>149</td>
<td>Rated</td>
<td>92</td>
</tr>
<tr>
<td>Comet</td>
<td>73</td>
<td>Gauze</td>
<td>65</td>
<td>Merit</td>
<td>71</td>
<td>Rebel</td>
<td>146</td>
</tr>
<tr>
<td>Crawl</td>
<td>101</td>
<td>Genesis</td>
<td>112</td>
<td>Metal</td>
<td>126</td>
<td>Rest</td>
<td>119</td>
</tr>
<tr>
<td>Crust</td>
<td>76</td>
<td>Gloom</td>
<td>88</td>
<td>Miser</td>
<td>82</td>
<td>Rip</td>
<td>106</td>
</tr>
<tr>
<td>Cupid</td>
<td>183</td>
<td>Gnash</td>
<td>171</td>
<td>Month</td>
<td>160</td>
<td>Reign</td>
<td>23</td>
</tr>
<tr>
<td>Curve</td>
<td>50</td>
<td>Gnome</td>
<td>39</td>
<td>Mouse</td>
<td>41</td>
<td>Rifle</td>
<td>106</td>
</tr>
<tr>
<td>Date</td>
<td>125</td>
<td>Guest</td>
<td>24</td>
<td>Mural</td>
<td>7</td>
<td>Rivet</td>
<td>124</td>
</tr>
<tr>
<td>Delta</td>
<td>63</td>
<td>Guide</td>
<td>187</td>
<td>Music</td>
<td>132</td>
<td>Roast</td>
<td>9000</td>
</tr>
<tr>
<td>Depot</td>
<td>148</td>
<td>Gunny</td>
<td>46</td>
<td>Robin</td>
<td>74</td>
<td>T</td>
<td></td>
</tr>
</tbody>
</table>

### CAUTION

- Personal Identification Waived

### VIGILANCE

- Insert “CAU” after the money word or words in the money order message

- Insert “VIG” after the money word or words in the money order message

---

**Note:** The diagram and text provide instructions on how to encode and decode messages using a two-part code system.
Comprehension
Comprehension

• “A code reflects the world at a particular instant, and as the world moves on it outmodes the code. New products, new ways of doing things, new political or economic facts begin to make its vocabulary old-fashioned.” (Kahn)

• Code books present a picture of a given era

• Code books could also be used for translation
A Bygone Age

• “Marriage has been arranged between _____” (*Unicode*, 1897)

• “Will lunch with you today” (*Unicode*)

• “Roman Catholic intrigue” (*China Inland Mission Private Telegraph Code*, 1907)

• “Send women on shore to wash” (*Popham’s Naval Signal Code*, under “Military and Technical Terms”, 1816)

• Professions: “Castle-keeper” (*International Police Telegraph Code*, 1930)
Cultural Norms: An Illuminated Persian Government Codebook (1901)
Sending Chinese Characters

• 4-digit/3-letter link-layer encoding for each Chinese character

• Widely used in China until about 10 years ago — faxes and cell phones have taken over

• Code points are still used today for names on official forms: dialect-independent, unambiguous, etc.
Copyright Infringement

• The U.S. hadn’t signed the Berne Convention; books weren’t protected here unless printed here first. Some code books were widely pirated.

♦ British publishers sometimes printed the first edition in the U.S. to avoid that

♦ Note — pirate editions couldn’t be imported into the British empire

• The code words themselves were valuable; those were pirated, too
Summary
Most Functions Existed at All Layers

<table>
<thead>
<tr>
<th></th>
<th>Link</th>
<th>Codebook</th>
<th>Plaintext</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Compression</strong></td>
<td>Morse code elements</td>
<td>Careful phrase selection</td>
<td>Restricted word choice; sentence fragments</td>
</tr>
<tr>
<td><strong>Correction</strong></td>
<td>Different links have different error properties; read-back</td>
<td>Mutilation tables; terminal indices; Hamming distance; check digits</td>
<td>Use code words for numbers</td>
</tr>
<tr>
<td><strong>Confidentiality</strong></td>
<td>Avoid exposed links (i.e., radio; other countries’ wires)</td>
<td>Superencryptions; secret codebooks</td>
<td>Semantically combine fields</td>
</tr>
</tbody>
</table>

There were generally tradeoffs in convenience, performance, efficiency, etc.
Parting Thoughts

• Telegraph codebooks were used in Australia until at least 1972, and in China until around 2000

• What we do today is the evolution of what was done then

• Huffman didn't invent compression; Hamming didn't invent error correction; NIST didn't invent encryption

• (Draft paper at papers/codebooks.pdf on my web page.)
Acknowledgments

• Jim Reeds

• Hang Zhao, Seung Geol Choi, Evelyn Guzman, Malek Ben Salem, Arezu Moghadam, and Ted Lemon

• The research libraries and librarians of the world

• Google Books
Compression, Correction, Confidentiality, and Comprehension

Steven M. Bellovin
smb@cs.columbia.edu
http://www.cs.columbia.edu/~smb

+1 212-939-7149
Department of Computer Science
Columbia University