Fred Cohen & Associates - Analyst Report and Newsletter

Welcome to our Analyst Report and Newsletter

Book code cryptography may be nearly dead

Book codes were historically one of the toughest sorts of cryptographic systems to break if well used. The basic notion is that you and I share the same book and write messages by encoding {page, line, word} in messages. If I send you a {27, 13,4}, {23,3,8}, etc. you can read it, but nobody else will be able to unless (1) they have the same book code and use it, (2) we use the same codings (or ones close to the same places in the book) more than once, or (3) we use words near the beginnings or endings of sentences, paragraphs, chapters, etc. The reason for (2) is that they might be able to determine valid sentences from words coded too close to each other, or use the reuse of words to reduce message uncertainty. The reason for (3) is that words starting and ending sentences, etc. are statistically different from others.

So what has changed?

You can imagine that with modern technology, I could get enormous book codes, automate the selection of word locations with pseudo-random selection from the set of all locations of each word desired, and do the coding more rapidly than most standard cryptographic systems do it. Indeed it's trivial to implement such a system reasonably well. As we use it more and more, we will run out of space in the book code that is unused, but it's easy enough to go to the next book (we have to share the secret of which book of course, but we will do that at a personal meeting somewhere). That makes it far easier to use, but there is a problem...

Suppose I am Google and I have the text to essentially every book on the market as well as papers modern and historical, and on and on. Suppose I make an index of all the words in every such book, keying them to the locations in each of the works. If I now reverse the index, I go from a translation from each word into all of the different possible codings for each book into a mapping from each possible coding into all of the {word, book} pairs. At this point, it is game over for book codes if Google has the books. All they do is take the codings of your sentences, map them into all possible word sequences for each of the billion or so books, and use a syntax checker (a natural language parser and verifier such as those used for speech to text programs) to eliminate all of the books that don't produce valid sentence syntaxes (in all languages all at once by the way). As I continue to get messages, I rapidly reduce the set of possible book codes until I get the one and only book that makes sense of it all. And at that point it is game over. And it can be fully automated, very fast (indices are linear time and space in the length of all books, lookups are constant time and linear space with the total size of the book codes, and graph analysis against valid syntax is also linear time in the size of the linguistic graph (I think).

Conclusions

I think it is game over for book codes based on widely available text, and those who have long thought this to be a secure method may soon find that even ancient messages not previously decoded will fall to modern technology.

