Selective Encryption of Text Files with Huffman Coding

Van T. Phu (New Jersey Institute of Technology), Advisor: Tom Lookabaugh (University of Colorado at Boulder).

Selective encryption is the technique of encrypting some parts of a compressed data file while leaving others unencrypted. Selective encryption is not a new idea. It has been proposed in several applications, especially in the commercial multimedia industry. However, selective encryption of losslessly compressed text files has not been explored, and that is the focus of our research. Through the study, we carefully studied how selective encryption can achieve a high level of effectiveness. By this, we mean a strategy in which even a small fraction of encrypted bits can cause a high ratio of damage to a file if an attacker attempts to decode it without decrypting the secured portions.

In this study, we combined the encrypting and compressing processes to consider the choices of which types of bits are most effective in the selective encryption sense when they are changed. And so, instead of encrypting the whole file bit by bit, we changed only these highly sensitive bits. Moreover, by combining the compression and encryption tasks and reducing the total encryption work required, we can achieve a savings in system complexity.

We used Huffman coding as the compression scheme for the text files. To measure the damage inflicted on a text file when it is decoded without decryption, we used the Levenshtein distance (D_{SID}) – the minimum number of substitution, insertion and deletion operations that are needed to make two files identical.

Experiments were carried out with some simple cases and some particular text files. In the simple cases, the results diverged from our expectations due to the complexities in the alignments of characters when calculating D_{SID}. The results from our experiments with some real text files, however, were very encouraging. With a ratio of encryption of only 10 to 20 %, we observed damage to an attacker's unauthorized decoding on the order of 70 – 85 % in 10 different text files that were picked randomly.