10
Conclusão e Trabalhos Futuros

Este trabalho teve início com a necessidade de se prover segurança para coleções textuais comprimidas. Soluções serializadas, onde primeiro o texto é comprimido e depois cifrado, e então posteriormente decifrado e descomprimido, são atualmente rápidas e eficientes, mas não permitem que as funções de recuperação de informações (RI), como busca e indexação sejam mantidas. Nesse caso, é necessário que o texto seja totalmente decodificado para que o acesso aleatório à informação seja possível.

O foco deste trabalho então foi a criação de algoritmos cripto-compressores. Foi definido neste trabalho o termo cripto-compressor para referenciar algoritmos que implementem as funções de cifragem e compressão de dados simultânemente. E nesse trabalho, que mantenham intactas as funções de RI. Além disso, podem ser assimétricos no desempenho, isto é, o importante é a rápida descompressão, podendo a compressão ser feita apenas uma vez pelo dono da coleção, sem tanto compromisso de desempenho. O objetivo é criar algoritmos cripto-compressores que se baseiem em algoritmos conhecidos de compressão de dados, sobre os quais se embute a segurança. Assim, somente pessoas autorizadas, que detenham uma chave secreta, têm acesso a coleções de dados, por exemplo em CD, e que então possam fazer buscas e acessar aleatoriamente partes do texto.

Primeiramente, alguns algoritmos de compressão de dados foram analisados, como os códigos de Huffman, os baseados no run-length, os códigos aritméticos, a transformada de Burrows-Wheeler (BWT), entre outros. O algoritmo escolhido foi o de Huffman, devido a expertise sobre esse algoritmo que o nosso grupo de pesquisa possui, e devido ao seu poder de sincronização e de suporte a funções de RI. A maior desvantagem do Huffman é ser um método semi-estático, ou seja, necessita de duas varreduras sobre o texto: uma para o cálculo das probabilidades e então outra para a codificação. Métodos adaptativos, que só necessitam de apenas uma passagem pelo texto, são geralmente rápidos e eficientes, mas tem a desvantagem de que a descompressão deve sempre iniciar do início e nenhum acesso aleatório é possível.

E então, foram analisados várias técnicas criptográficas e como elas

poderiam ser aplicadas para incluir segurança em algoritmos de compressão de dados. Foram analisadas diversos sistemas criptográficos simétricos (3DES, o AES, etc.), de chave pública (RSA, etc.), de resumo da mensagem (MD5, SHA1, etc.), arquiteturas como redes de Feistel, S-Boxes, permutação, sistemas clássicos, homofonia, esteganografia, entre outros, de tal forma a dar subsídios à criação de sistemas criptográficos próprios. As técnicas de criptoanálise diferencial, linear, e redução matemática foram estudadas também em detalhes para permitir que fossem feitas auto-análises das propostas de segurança.

Foram então feitas várias tentativas de criação de algoritmos cripto-compressores, e os mais promissores foram registradas nesta tese, sendo baseados principamente nos códigos de Huffman canônicos com varredura de palavras e na técnica de substituição homofônica. Para a avaliação da segurança dos métodos foram feitos vários experimentos práticos e argumentações teóricas do sigilo, como através de redução matemática de problemas NP-Completos. A ferramenta do NIST ("A statistical test suite for random and pseudorandom number generators for cryptographic applications") também foi muito útil em alguns casos.

No estudo de pesquisas correlatas, poucos artigos foram encontrados. Não há notícias de algum algoritmo cripto-compressor sendo explorado comercialmente. Sistemas operacionais como o Windows ou ferramentas populares como o WinZIP adotam soluções serializadas, e então não permitem, por exemplo, busca nos arquivos comprimidos-cifrados.

De modo geral, os algoritmos apresentam pequena expansão no tamanho se comparados com o algoritmo de compressão original, pequeno overhead no desempenho de compressão e ótima velocidade de descompressão, sendo assim voltados para a aplicação a que se propõe.

Foram propostos 4 algoritmos:

ADDNULLS Uso da técnica de estaganografia com a inserção seletiva de símbolos nulos para ocultar a mensagem comprimida dentro de falsos bits. O método é formalizado e a expansão resultante da segurança é calculada. O algoritmo foi implementado e apresentou bom desempenho e pouca expansão.

HHC Uso de códigos de Huffman canônicos, substituição homofônica e distribuição diádica para criar um algoritmo que embaralha a mensagem final através de uma permutação dos símbolos homofônicos em cada nível da árvore canônica de Huffman. Vários experimentos práticos foram implementados e documentados.

RHUFF Uso de uma permutação inicial do texto de entrada e a quebra em blocos de tal forma a criar textos cujos tamanhos sejam potências de dois. A

expansão é conseguida pela inserção de símbolos falsos somente no último e menor bloco. O algoritmo gera um fluxo de bits aleatórios. Pode ser utilizado para a aleatorização da entrada em um outro algoritmo.

HSPC2 Técnica para inserção de segurança em códigos de prefixo. Pode usar o [RHUFF] para aleatorizar a entrada. Argumentação teórica baseada na redução matemática ao problema da mochila (SUS).

Este trabalho pode ser estendido através da criação de novos algoritmos, utilizando as mesmas técnicas descritas nessa tese, ou mesmo com algoritmos de compressão promissores como a BWT. O ideal é que experimentos práticos e argumentações teóricas devem ser feitas para cada algoritmo, ou mesmo para cada módulo do algoritmo, o que não foi feito totalmente para os 4 algoritmos aqui propostos, mas seria deséjável. Além disso, a maioria das técnicas de criptoanálise voltadas para sistemas criptográficos simétricos não se aplicam bem aos algoritmos propostos. Seria interessante a definição de técnicas de criptoanálise para problemas desse tipo.

Alguns problemas correlatos também não foram explorados em detalhes, mas poderiam ser mais desenvolvidos como a codificação otimizada do modelo, o uso de algoritmos de árvores esqueleto (17) para a aceleração da descompressão, e técnicas para redução do modelo (35).