Um método simples para detecção on-the-fly de arquivos e suas mutações aplicado ao combate à pedofilia e outros crimes na Internet

Breno Rangel Borges Marchetti, *Perito Criminal Federal, Departamento de Polícia Federal do Brasil*

Abstract— A eficácia no combate a crimes na Internet como a pedofilia, captura de senhas bancárias e divulgação de informações sigilosas está diretamente relacionada aos processos de obtenção e análise de dados em tempo hábil para que as medidas repressivas necessárias sejam tomadas. Infelizmente, estes dois processos tendem a serem distintos e freqüentemente demorados. Não bastasse a volatilidade dos arquivos e códigos maliciosos que trafegarem pela rede, há freqüente alteração no conteúdo destes, o que prejudica a eficácia de diversas técnicas difundidas atualmente, como por exemplo, detecção baseada em assinaturas e resumos. Neste contexto, é proposta uma metodologia que alia extração e seleção de arquivos on-the-fly utilizando a técnica NCD (Normalized Compression Distance). Esta técnica explora algoritmos de compressão de arquivos de forma a detectar similaridades entre quaisquer tipos de dados de interesse. Como prova de conceito esta metodologia foi implementada e os resultados são apresentados e discutidos.

***Index Terms*— Crimes, Internet, pedofilia, NCD.**

I. INTRODUÇÃO

A Internet por sua natureza dinâmica e pseudo-anônima é um meio que favorece a prática de vários crimes cuja materialização é viabilizada pela transferência de dados e arquivos sob os mais diversos artifícios.

A captura e o reconhecimento destes dados são peça chave em processos preventivos ou investigativos que não raramente encontram entraves de natureza técnica.

O combate a estes crimes é ainda dificultado pela natureza volátil dos dados ao trafegarem pela rede assim como a freqüente modificação (mutabilidade) destes dados de interesse, doravante chamados de *alvos*. Desta forma, uma vez estabelecidos os *alvos* a captura do tráfego de rede sob suspeita deve ser feita no menor tempo possível assim como a posterior extração e o reconhecimento dos *alvos* e/ou suas variações.

Captura, extração e seleção - estas três etapas devem ser realizadas preferencialmente de maneira simples e transparente para o analista ou investigador, o qual geralmente não possui o conhecimento técnico necessário para a captura e extração dos dados, limitando-se a analisar e selecionar os dados ora extraídos. Esta fase de seleção e análise dos dados freqüentemente demanda tempo considerável, fato que muitas vezes prejudica o processo investigativo/repressivo onde a oportunidade de ação é fundamental e não pode ser perdida.

Crimes na Internet como a pedofilia e a difusão de códigos maliciosos para captura de senhas bancárias são facilitados pela dificuldade de detecção da mutabilidade de arquivos e códigos executáveis.

No caso específico do crime de pedofilia, há imagens similares onde o aspecto pictório é mantido mesmo quando a imagem original é editada, como por exemplo, no caso de aplicação de efeitos nas cores, inserção de textos, omissão de trechos ou redimensionamento. No caso de mutabilidade de códigos executáveis maliciosos, pode haver diferentes rotinas com o mesmo objetivo ou finalidade.

É fácil perceber então que a mutabilidade dos dados pode facilmente prejudicar as técnicas atualmente difundidas de individualização de arquivos baseado em resumos criptográficos (hash) ou mesmo de seleção baseada em assinaturas.

Neste cenário, é proposta uma metodologia cujo objetivo é a seleção automatizada de arquivos que apresentam certo grau de similaridade com aqueles fornecidos como referência, através da abordagem *NCD* (Normalized Compression Distance). Esta abordagem utiliza compressão para detectar similaridade entre dois arquivos.

A metodologia proposta pode ser utilizada para detectar similaridades em quaisquer tipos de arquivo, porém este artigo focará arquivos do tipo imagem, muito utilizadas por pedófilos que encontram na Internet local propício para a divulgação deste tipo de material.

II. METODOLOGIA

A metodologia proposta tenta unir e automatizar as fases de captura de dados, extração e seleção de dados relevantes de modo a facilitar o trabalho de análise pelo analista/investigador. Naturalmente, será utilizado como referência o protocolo TCP/IP devido a este ser o padrão *de facto* utilizado em comunicações de dados com controle de correção de erro através da Internet.

No protocolo TCP/IP, os dados de interesse estão contidos em vários pacotes de dados que podem chegar ao destinatário fora de seqüência e conter dados associados a protocolos utilizados pelas aplicações dos usuários. Está fora do escopo deste artigo a discussão técnica sobre a remontagem destes pacotes e extração dos arquivos ali presentes, visto que atualmente existem disponíveis várias ferramentas *Open Source* que desempenham tais funções.

A. *Extração de arquivos sob demanda*

A ferramenta escolhida foi a *tcpxtrac*t, disponível no sítio *tcpxtract.sourceforge.net*, que é uma ferramenta *Open Source*, de uso gratuito e extrai arquivos do tráfego TCP/IP baseando-se em assinaturas de cabeçalho ou rodapé (técnica usualmente chamada de *"carving"*).

Esta ferramenta possibilita a realização da extração de mais de 26 tipos de arquivos popularmente utilizados, entre eles figuras *JPEG,BITMAP, GIF, PNG*, documentos do Microsoft Word e até mesmo arquivos compactados *ZIP*, dentre outros. Além de realizar a captura *on-the-fly* do tráfego TCP/IP, apresenta outro recurso importante que é a possibilidade de extração de arquivos presentes em pacotes previamente capturados armazenados no popular formato *tcpdump*, abrindo assim um leque de ferramentas para ser utilizada na tarefa específica de captura dos pacotes.

B. *Métrica de Similaridade NCD*

Uma vez estabelecida a maneira como os arquivos serão recuperados a partir do tráfego TCP/IP da Internet, é aplicada a métrica de similaridade *NCD* que tem como objetivo classificar o arquivo extraído em similar ou não em comparação a outros previamente especificados pelo analista ou investigador.

A métrica *NCD* (Normalized Compression Distance) utiliza algoritmos de compressão de dados para inferir similaridade entre dois arquivos. Técnicas de compressão de dados basicamente procuram blocos de código repetido substituindo estes por um código definido. Algoritmos de compressão mais sofisticados procuram padrões de diversos bytes realizando o mesmo tipo de substituição. O processo também pode ser aplicado recursivamente obtendo-se assim taxas de compressão ainda maiores.

Considerando que arquivos muito semelhantes conterão vários padrões de dados semelhantes, um bom algoritmo de compressão deve comprimir dois arquivos semelhantes juntos obtendo no final um arquivo com o tamanho bem próximo ao de apenas um deles compactado separadamente. A fórmula para a métrica *NCD* pode ser definida como:

$$NCD = \frac{cmprd(a1 + a2) - \min(comprd(a1), comprd(a2))}{\max(\text{cmprd}(a1), \text{cmprd}(a2))}$$

Na fórmula acima, *NCD* é um coeficiente que varia de 0 a 1, *a1* e *a2* são os arquivos a serem comprimidos, *cmprd* é uma função de compressão que gera uma arquivo comprimido a partir de *a1* e/ou *a2*, retornando o tamanho do arquivo gerado. *Max* e *Min* são funções que retornam o tamanho máximo e mínimo respectivamente a partir de *a1* e *a2*.

O coeficiente *NCD* apresentará valores bem próximos a 1 para arquivos totalmente diferentes entre si, e valores próximos a 0 para arquivos similares. Valores intermediários podem indicar o grau de similaridade entre dois arquivos. Este coeficiente pode ser então utilizado no critério de seleção dos arquivos que são extraídos a partir do tráfego TCP/IP conforme explanado anteriormente.

Um bom algoritmo de compressão a ser utilizado em casos práticos deve apresentar elevada taxa de compressão para que a métrica *NCD* seja eficaz.

A seleção do método de compressão baseou-se em algoritmos largamente utilizados e implementados como o *BZip2*, *ZIP*, *PPMd e LZMA* (*Lempel-Ziv-Markov chain-Algorithm*). Alguns utilitários de uso gratuito que implementam esses algoritmos são o *bzip2*, *gzip (ZIP), e 7za (LZMA)*. Foi realizado dois testes para escolher o utilitário para ser utilizado na implementação da métrica *NCD*. O primeiro teste consistiu em compactar um diretório contendo arquivos tipo texto referentes ao código fonte do kernel do Linux versão 2.6.25.9 que no total somavam 308 MB. Os resultados obtidos são listados na *tabela 1*.

Tabela 1 – Teste de compactação sobre arquivos texto.

Utilitário	Tempo de compressão	Tamanho compactado	Tamanho em relação ao arquivo original
BZIP2	1m 27s	47 MB	15%
GZIP	44s	60 MB	19%
7za (LZMA)	7m 47s	40 MB	**12%**

O utilitário 7za que utilizou o algoritmo LZMA foi o que apresentou a melhor taxa de compressão, no entanto apresentou o pior tempo de compressão. Foi realizado um segundo teste com uma imagem no formato BITMAP de 2.1 MB (*figura 1*) contendo diversas cores e os resultados são mostrados na *tabela 2*.

Figura 1 – Imagem BITMAP utilizada no teste de compressão.

Tabela 2 – Teste de compressão sobre imagem BITMAP

Utilitário	Tempo de compressão	Tamanho comprimido	Tamanho em relação ao arquivo original
BZIP2	0.5 s	1011KB	48%
GZIP	0.2 s	1.2 MB	57%
7za (LZMA)	1.6 s	969 KB	**46%**

Um terceiro teste foi realizado com arquivos executáveis, e novamente o utilitário 7za que implementa o algoritmo LZMA obteve a melhor taxa de compressão.

A eficácia da técnica *NCD* está diretamente relacionada a uma boa taxa de compressão, então o utilitário 7za utilizando o algoritmo LZMA foi o escolhido para o teste desta métrica nos arquivos extraídos do tráfego TCP/IP.

III. IMPLEMENTAÇÃO DA METODOLOGIA

Uma vez definida as ferramentas a serem utilizadas, foi implementado um software na linguagem *C* através do compilador *GCC* para Linux, que aceita como entrada um conjunto de arquivos escolhidos pelo usuário e utiliza este conjunto como referência para detecção de possíveis arquivos similares dentre os arquivos extraídos do tráfego TCP/IP conforme explanado anteriormente.

Um arquivo é considerado então similar caso esteja abaixo do valor limite definido para o coeficiente *NCD*.

O software desenvolvido monitora o diretório para onde são extraídos os arquivos a partir do tráfego TCP/IP, ou a partir de outra fonte qualquer, aplicando a métrica *NCD* sempre que existam arquivos em tal diretório. Caso o coeficiente *NCD* obtido esteja abaixo do limite definido, o arquivo é considerado como similar e é armazenado, caso contrário, o arquivo é descartado.

A implementação da metodologia é mostrada na *figura 2*.

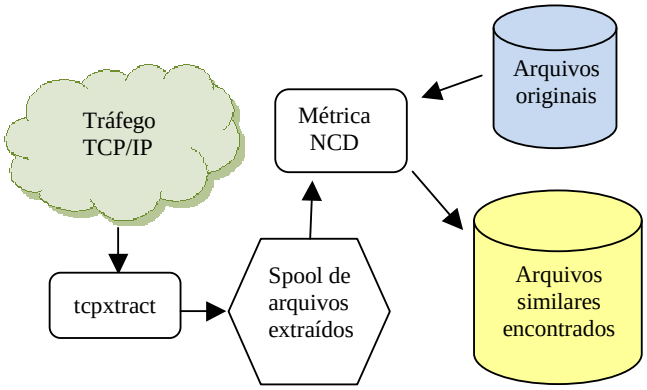

Figura 2 – Implementação do reconhecimento de arquivos similares sob demanda (on-the-fly).

É importante notar aqui, que a métrica *NCD* não funciona bem para casos em que os arquivos já se encontram compactados, como é o caso de arquivos do tipo *ZIP*, figuras *JPEG*, sons no formato *MP3* e executáveis no formato *UPX*, muito encontrados na Internet. Um método simples para contornar este problema é a descompactação destes arquivos antes da aplicação da métrica.

Em casos de pedofilia, há uso freqüente de arquivos do tipo JPEG, que muitas vezes são gerados na própria câmera fotográfica utilizada pelo pedófilo.

Arquivos texto são conhecidos por apresentarem excelente taxa de compressibilidade. Um método experimental utilizado foi a transformação das imagens *JPEG* em figuras compostas por caracteres *ASCII* que tenta manter o aspecto pictório da imagem original utilizando caracteres *ASCII*.

Vários programas convertem uma imagem para caracteres texto através do método de quantização de vetorial. Este método re-amostra a imagem em uma escala de cinza com menos de 8-bits de precisão, e depois associa um caractere ASCII para cada valor.

Nos testes foi convencionado que as imagens de saída teriam sempre a mesma largura e altura. Uma das vantagens da conversão pra texto é o pequeno tamanho do arquivo resultante e a conseqüente rapidez de compactação do mesmo.

Figura 3 – Aspecto *da figura 1* em caracteres ASCII

IV. TESTE DA MÉTRICA NCD

O teste foi conduzido com imagens no formato *JPEG*, obtidas na Internet. As imagens originais foram então editadas e enviadas para uma conta de webmail. Posteriormente, o teste foi iniciado fornecendo as imagens originais como referência para o software desenvolvido que implementa a métrica *NCD*.

O teste consistiu em acessar na Internet as imagens originais e suas modificações que foram enviadas para um webmail.

Toda a navegação na Internet foi feita monitorada pelo software *tcpxtract*, que foi capaz de extrair as imagens do tráfego TCP/IP sob demanda (*on-the-fly*). A métrica *NCD* era então imediatamente aplicada tão cedo os arquivos eram extraídos do tráfego TCP/IP e armazenados no diretório monitorado.

Os arquivos modificados que foram encontrados e selecionados automaticamente são mostrados na *tabela 03*.

As imagens encontradas e selecionadas automaticamente no teste são variantes da figura original que foi editada com efeitos como redimensionamento, adição de textos, alteração das cores, omissão ou acréscimo de trechos.

O coeficiente limite *NCD* utilizado no teste foi de 0.4, sendo que neste caso observou-se raríssimos casos de falso positivo na seleção dos arquivos.

Também foram realizados testes semelhantes de modificação em outros tipos de arquivo como documentos do Microsoft Word, planilhas Microsoft Excel, arquivos texto, executáveis do Microsoft Windows e os resultados foram ainda mais animadores.

Como era de se esperar, o teste realizado apresentou como limitador de velocidade o tempo de compressão dos arquivos pelo algoritmo empregado, neste caso o LZMA. O teste foi realizado em um link *ADSL* de 2MB/s e a estação de análise utilizou um processador da Intel Core 2 Duo de clock 2 GHz.

A navegação foi realizada de forma a simular uma navegação cotidiana na Internet, realizando-se buscas por

imagens no sítio http://images.google.com, assim como outros sítios escolhidos aleatoriamente. Esta configuração diante de uma navegação feita normalmente por usuários da Internet era capaz de fornecer centenas de figuras em menos de um minuto.

O gargalo observado não foi suficiente para prejudicar a velocidade de extração e seleção das imagens sob demanda.

Tabela 3 – Imagens modificadas encontradas a partir da imagem original.

Imagem original	Imagens encontradas		
			
			
			

V. Conclusão

A métrica *NCD* apresentou resultados satisfatórios que possibilitam a detecção de similaridade nos mais diversos tipos de arquivos, abrindo um leque para as mais diversas aplicações em casos práticos. A detecção sob demanda de arquivos e suas variações foi feita com sucesso e a métrica *NCD* apresentou vantagens como a capacidade de ser aplicada em virtualmente qualquer tipo de arquivo, a facilidade de implementação e taxas animadoras de acerto.

Esta técnica de background teórico simples tem sua eficiência limitada pelo tempo gasto na compressão e na taxa de compressibilidade intrínsecos ao algoritmo de compactação utilizado.

Geralmente, quanto melhor a taxa de compressão maior será o tempo gasto nas fases de compressão utilizadas pela métrica *NCD*. Felizmente, as rotinas de compactação são facilmente paralelizáveis, fato que pode tirar proveito dos processadores multi-núcleo atualmente em constante desenvolvimento e já largamente disponíveis no mercado a preços acessíveis.

Rotinas que tiram proveito de execução em paralelo podem ser facilmente desenvolvidas e implementadas através de bibliotecas especializadas na geração deste tipo de código para os atuais processadores multi-núcleo, como por exemplo a biblioteca *Thread Building Blocks* desenvolvida inicialmente pela Intel e agora Open Source.

A taxa de compressão pode ainda ser melhorada através da aplicação de um pré-processamento específico para o tipo de arquivo em questão, com o objetivo de aumentar a quantidade de padrões presentes no arquivo e conseqüentemente aumentando a eficiência do algoritmo de compressão utilizado pela métrica. Um exemplo é a descompressão de tipos de arquivo que já utilizam algum tipo de compressão, como arquivos tipo *JPEG,PNG,ZIP,BZ2,MP3*, executáveis compactados com *UPX*, arquivos de vídeo em geral, dentre outros.

Arquivos de vídeo são um caso especial onde dados de imagem e áudio são armazenados juntos mas podem ser extraídos e analisados separadamente para a detecção de similaridade através desta métrica.

A métrica *NCD* também pode ser utilizada para tentar classificar códigos executáveis desconhecidos na análise de *malwares*, *bankers* e outros tipos de códigos executáveis maliciosos, área em que alguns experimentos já foram realizados com sucesso por empresas fabricantes de antivírus.

Diante dos fatos e resultados, a metodologia proposta pode ser mais uma simples e importante ferramenta para auxiliar no combate a diversos crimes cometidos através da Internet.

REFERÊNCIAS

[1] M. Li, X. Chen, X. Li, B. Ma, P.M.B. Vitanyi. The similarity metric, IEEE Trans. Inform. Th., 50:12(2004), 3250- 3264.

[2] M. Li and P.M.B. Vitanyi. An Introduction to Kolmogorov Complexity and its Applications, Springer-Verlag, New York, 2nd Edition, 1997.