PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS

Bacharelado em Ciência da Computação

Lara Brígida Rezende Souza
Matheus Moreira Sorrentino
Raul da Cruz Fonseca
Victor Cabral de Souza Oliveira

DETECÇÕES DE ANOMALIAS EM ARQUIVOS CRIPTOGRAFADOS

Belo Horizonte

2023

Lara Brígida Rezende Souza
Matheus Moreira Sorrentino
Raul da Cruz Fonseca
Victor Cabral de Souza Oliveira

DETECÇÕES DE ANOMALIAS EM ARQUIVOS CRIPTOGRAFADOS

Projeto de Pesquisa apresentado na disciplina Trabalho Interdisciplinar III - Pesquisa Aplicada do curso de Ciência da Computação da Pontifícia Universidade Católica de Minas Gerais.

Belo Horizonte

2023

RESUMO

Nesta pesquisa os autores irão abordar o tema de detecção de anomalias em arquivos criptografados, trazendo os problemas principais, os objetivos e a revisão bibliográfica como grandes enfoques.

A revisão bibliográfica contará com sete artigos relacionados ao tema e os autores trarão a correlação daquilo que está sendo dito nos artigos com o tema abordado por eles nesta pesquisa.

Palavras-chave: detecção; anomalias; arquivos criptografados; pesquisa; revisão bibliográfica; objetivos; problemas.

SUMÁRIO

1 INTRODUÇÃO

A detecção de anomalias em arquivos criptografados é fundamentada na identificação de caracteres potencialmente discrepantes em relação ao conteúdo geral do arquivo, o que pode resultar em desafios na interpretação do documento ou na ocorrência de problemas relacionados à integridade dos arquivos. Tais anomalias podem ter origem em diversas fontes, tais como erros na codificação, falhas na transmissão de dados ou tentativas de acesso não autorizado.

A identificação atempada de anomalias nesses arquivos desempenha um papel crucial na prevenção de falhas na interpretação ou corrupção dos dados, bem como na mitigação de possíveis ameaças à segurança da informação. Neste contexto, o problema de detecção de anomalias implica principalmente na identificação de caracteres que fogem do padrão em relação ao conteúdo do arquivo.

Essas anomalias são comumente identificadas após tentativas infrutíferas de codificação e transmissão de dados, resultando na corrupção de arquivos e/ou na perda de informações. Esta questão reveste-se de significativa importância, notadamente em virtude da natureza crítica dos dados, atualmente considerados um ativo extremamente valioso. Em um mundo globalmente interconectado, onde a troca contínua de informações é imperativa, qualquer lacuna nas informações pode potencialmente ter impactos sistêmicos abrangentes

Este trabalho está organizado da seguinte forma. A seção 1.1 representa os objetivos procurados na pesquisa. O capítulo 2 apresenta o referencial teórico usado neste trabalho.

1.1 Objetivos

O objetivo geral deste projeto é apresentar uma análise abrangente das diversas abordagens para a detecção de anomalias, com um foco específico em arquivos criptografados, que foram previamente propostas. Além disso, buscaremos investigar o impacto das anomalias nesses arquivos e compreender as potenciais consequências resultantes de sua presença.

1.1.1 Objetivos específicos

Os objetivos específicos deste projeto são:

1. **Identificação de Outliers:**

   A identificação de outliers está diretamente relacionada aos arquivos criptografados, pois busca desenvolver uma técnica específica para identificar dados destoantes dentro desses arquivos que podem indicar a presença de anomalias.

2. **Análise de Custos e Recursos:**

   A análise de custos e recursos é fundamental quando se trata de arquivos criptografados, uma vez que esses arquivos são muitas vezes utilizados para armazenar informações sensíveis e a análise de anomalias pode ser uma ação custosa.

3. **Avaliação de Desempenho:**

   A avaliação de desempenho e complexidade desempenha um papel fundamental no contexto de arquivos criptografados, uma vez que esses arquivos são frequentemente utilizados para armazenar informações altamente sensíveis. A detecção de anomalias pode envolver processos que consomem recursos substanciais, tornando essencial uma análise minuciosa dos custos envolvidos.

2 REVISÃO BIBLIOGRÁFICA

Este capítulo apresenta os artigos que os autores utilizaram como referência para realizar a pesquisa acerca do tema de detecção de anomalias em arquivos criptografados. Ao todo foram utilizados sete artigos que a seguir serão correlacionados com o tema abordado na pesquisa.

Os autores Provotar, Linder e Veres (2019) tratam do problema de utilizar algoritmos padrões de detecção de anomalias ao se trabalhar com a mineração de dados, devido à natureza desconhecida que pode afetar o todo. Sendo assim, foi proposto o uso de uma rede neural que aprende um encordoamento de um conjunto de dados, ignorando o noise. O artigo e a pesquisa se correlacionam na perspectiva de teste e amostragem de formas diferentes de tratamento de dados e detecção de anomalias.

Nesse artigo, os autores Bursic, Cuculo e D'Amelio (2020) abordam o crescimento da complexidade dos sistemas de computador, ao ponto de que a inspeção manual do comportamento do sistema para fins de detecção de mau funcionamento se tornou inviável. Esse tipo de sistema depende de recursos feitos à mão, exige pré-processamento de log bruto e extração de recursos ou usam aprendizado supervisionado, necessitando de um conjunto de dados de log rotulado que nem sempre é facilmente obtido. A solução proposta foi um modelo de autoencoder profundo em duas partes com unidades de LSTM que não requer recursos artesanais, que gera uma pontuação de anomalia para cada entrada de log. O artigo e a pesquisa têm correlação no ponto de analisar anomalias em arquivos de dados, apesar da diferença no tipo de arquivo.

Os autores Landauer et al. (2018) discorrem, no artigo, sobre o problema de análise de log em grandes sistemas computacionais. Eles buscam tratar o problema da análise de log como um tema que necessita de uma análise mais dinâmica do problema, devido a possíveis empecilhos de segurança. A solução proposta para a problemática foi a criação de clusters de logs gerados em momentos próximos e avaliação de semelhanças entre os erros, buscando perceber uma brecha de segurança dentro do sistema de forma mais dinâmica, devido a maior facilidade de perceber um grande conjunto de logs gerados em um curto espaço de tempo como algo grande a ser averiguado e também propondo um sistema não supervisionado. O ponto em comum entre o artigo e o trabalho é o fato de

serem realizadas análises em uma base de dados para encontrar problemas e a tentativa de achar uma solução para evitar as anomalias.

A respeito das ideias de Maciąg et al. (2021) a descoberta não supervisionada de anomalias em dados de fluxo é um tópico de pesquisa com muitas aplicações práticas. Entretanto, um desafio é a coleta de dados de treinamento com anomalias rotuladas suficientes para aprendizagem supervisionada de um detector de anomalias, com o objetivo de implantá-lo posteriormente para identificar anomalias reais em dados de streaming. A resposta encontrada foi adaptar o classificador Online Evolution Spiking Neural Network (OeSNN) para a tarefa de detecção de anomalias. Os autores do artigo, como resultado, ofereceram um algoritmo de rede neural de pico em evolução on-line para detecção de anomalias não supervisionadas (OeSSN-UAD), que, ao contrário do OeSNN, funciona de forma não supervisionada e não separa os neurônios de saída em classes de decisão disjuntas. Correlacionando com o trabalho de pesquisa, o artigo também apresenta um problema relacionado a anomalias em dados, dada a diferença de que são dados para treinamento para aprendizagem supervisionada.

A partir da visão de Nedelkoski et al. (2020) a detecção de anomalias é muito importante na mineração de dados para atingir a segurança e confiabilidade em sistemas computacionais. De acordo com os autores, os logs são uma fonte de dados comum e importante para métodos de detecção de anomalias em quase todos os sistemas de computador. No artigo, o principal problema encontrado foi a limitação dos modelos de aprendizagem, por não serem capazes de aprender representações de log que descrevem as diferenças semânticas entre logs normais e de anomalias, ocasionando uma generalização pobre em logs não vistos. A solução encontrada é uma abordagem em que o conjunto de dados auxiliares seja suficientemente informativo para melhorar a representação dos dados normais, mas diversificado para regularizar contra overfitting e melhorar a generalização. A associação deste artigo com a pesquisa se dá no fato de anomalias causarem problemas na execução de tarefas, como no artigo, a dificuldade de modelos de aprendizagem com arquivos de log.

Segundo Zhu et al. (2018) muitas empresas migraram os dados para a nuvem utilizando serviços de sincronização e compartilhamento de arquivos (FSS), implatados para usuários móveis. Entretanto, isso trouxe um desafio, evitar que os usuários abusem do serviço FSS. Os autores do artigo abordam o problema através de um novo modelo de sistema que envolve detecção, rastreamento e revogação de anomalias. A solução encontrada por eles foi aplicar um novo sistema criptográfico baseado em chave pública de limite, chamada de criptografia hierárquica parcialmente ordenada (PHE). Assim como o trabalho de pesquisa, o artigo busca combater as anomalias que estão envolvidas com arquivos criptografados através de estratégias, no caso deste artigo, o sistema PHE.

De acordo com Xin et al. (2021) os problemas relacionados com segurança e vírus em terminais de dispositivos e os métodos de detecção de tráfego encriptado têm sido incapazes de satisfazer as necessidades de algumas empresas. No artigo é proposto um método de monitoramento de tráfego criptografado para usuários finais para realizar a detecção anormal de tráfego de usuários. Este artigo se relaciona com o tema do trabalho no quesito de que anomalias em arquivos criptografados geram problemas que afetam os usuários diariamente, no artigo citado, é especificado na área de tráfego de dispositivos, o que se relaciona com o trabalho de pesquisa, por se tratar de dados e a importância deles estarem intactos, sem que haja corrupção dos mesmos.

REFERÊNCIAS

BURSIC, S.; CUCULO, V.; D'AMELIO, A. Anomaly detection from log files using unsupervised deep learning. In: SEKERINSKI, E. et al. (Ed.). Formal Methods. FM 2019 International Workshops. Cham: Springer International Publishing, 2020. p. 200–207. ISBN 978-3-030-54994-7.

LANDAUER, M. et al. Dynamic log file analysis: An unsupervised cluster evolution approach for anomaly detection. Computers Security, v. 79, p. 94–116, 2018. ISSN 0167-4048. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167404818306333>.

MACIąG, P. S. et al. Unsupervised anomaly detection in stream data with online evolving spiking neural networks. Neural Networks, v. 139, p. 118–139, 2021. ISSN 0893-6080. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0893608021000599>.

NEDELKOSKI, S. et al. Self-attentive classification-based anomaly detection in unstructured logs. In: 2020 IEEE International Conference on Data Mining (ICDM). [S.l.: s.n.], 2020. p. 1196–1201.

PROVOTAR, O. I.; LINDER, Y. M.; VERES, M. M. Unsupervised anomaly detection in time series using lstm-based autoencoders. In: 2019 IEEE International Conference on Advanced Trends in Information Theory (ATIT). [S.l.: s.n.], 2019. p. 513–517.

XIN, G. et al. An anomaly detection method of encrypted traffic based on user behavior. In: Proceedings of the 2021 1st International Conference on Control and Intelligent Robotics. New York, NY, USA: Association for Computing Machinery, 2021. (ICCIR '21), p. 51–56. ISBN 9781450390231. Disponível em: <https://doi.org/10.1145/3473714.3473724>.

ZHU, Y. et al. Phe: An efficient traitor tracing and revocation for encrypted file syncing-and-sharing in cloud. IEEE Transactions on Cloud Computing, v. 6, n. 4, p. 1110–1124, 2018.