Uma Revisão Sistemática da Literatura em Forense Digital e Bancos de Dados

Ayane C. A. Fernandes 1; Kayque M. Siqueira 2, Michele A. Brandão 3; Danilo B. Seufitelli 4.

1 Ayane Cristina Alves Fernandes, Bolsista (CNPq), Técnico em Informática, Instituto Federal de Minas Gerais - campus Ribeirão das neves, Ribeirão das Neves - MG; ayanecristinamb@gmail.com
2 Kayque Meira Siqueira, Bolsista (CNPq), Técnico em Informática, Instituto Federal de Minas Gerais - campus Ribeirão das Neves, Ribeirão das Neves - MG; kayque.siq@gmail.com
3 Orientador: Michele A. Brandão, Pesquisador do IFMG, Campus Ribeirão das Neves; michele.brandao@ifmg.edu.br
4 Coorientador: Danilo B. Seufitelli, Doutorando UFMG, Departamento de Ciência da Computação; daniloboechat@dcc.ufmg.br

RESUMO

A popularização dos recursos digitais e o maior acesso a dados na Web possibilitou que crimes digitais fossem cada vez mais frequentes. A perícia forense é a ciência responsável por pesquisas, análises e investigações de evidências criminais em diversas áreas de perícia. No contexto dos crimes cibernéticos, o foco deste trabalho é utilizar a forense digital para investigar tais crimes no contexto de banco de dados. Desta forma, a forense digital objetiva identificar, preservar, recuperar, analisar e apresentar dados digitais durante uma investigação digital. Com o extensivo uso da Web e meios digitais, a forense digital tornou-se essencial, além de seu crescimento cada vez mais expressivo em pesquisas científicas. Portanto, este trabalho apresenta uma revisão sistemática da literatura sobre forense digital e bancos de dados. As publicações encontradas foram classificadas em quatro categorias principais, sendo elas: SGBD (Sistema de Gerenciamento de Banco de Dados), e SQLi (*SQL Injection*), *Data Building* e *Forensic Intelligence.* Para cada classificação, identificamos os principais problemas abordados na área, bem como direcionamos sobre oportunidades de pesquisa além de destacar as limitações nos trabalhos existentes.

INTRODUÇÃO:

A constante evolução do uso de dispositivos eletrônicos e sistemas computadorizados em diversos setores na contemporaneidade contribui para um aumento na ocorrência de diversos tipos de cibercrimes, por exemplo, a injeção de malwares, roubo de dados financeiros e ataques DDoS (Distributed Denial-of-Service) (Cruz e RODRIGUES 2018; JAQUES et al. 2018). Isso gera prejuízos materiais e imateriais aos alvos de tais ataques. Concomitantemente a esse fenômeno, o volume de dados e a quantidade de informações em trânsito no mundo digital também sofre um drástico aumento, viabilizando a existência de dados que podem ser utilizados na investigação forense desses crimes cibernéticos, atuando como evidências digitais.

A forense digital é uma ciência que além de ajudar na reconstrução de cibercrimes e no desenvolvimento de medidas de prevenção a tais eventos maliciosos, atua na busca, análise, identificação e categorização desse tipo de dado referenciado como evidência digital. Um exemplo da efetividade do uso da forense digital na resolução de crimes reais pode ser observado no trabalho de

Gonçalves et al. (2009), que descreve um caso real de investigação forense. É relatado que um profissional da área imobiliária foi a julgamento por um homicídio que havia cometido, sendo que uma das provas utilizadas na audiência foi a localização de seu dispositivo móvel no momento do crime. Na ocasião, foi constatado que o suspeito estava próximo ao local onde ocorreu o assassinato. A informação de rastreio do dispositivo foi extraída pela polícia a partir de técnicas utilizadas pela forense digital.

Entretanto, apesar da real importância de estudos relacionados à área da forense digital na resolução de crimes e investigações cibernéticas, ainda existem poucos trabalhos nesta área quando comparado a outros tópicos das diversas subáreas da Computação Aplicada. Faz-se então necessário o desenvolvimento de uma revisão sistemática sobre os trabalhos dessa área de estudo. De fato, um dos principais objetivos deste projeto é promover a melhor categorização, síntese e, consequentemente, facilitar a busca e o acesso a trabalhos sobre forense digital e bancos de dados.

METODOLOGIA:

Neste trabalho, a metodologia utilizada para a pesquisa se baseia na execução de 7 etapas adaptadas do protocolo de Kitchenham e Charters (2007) e são explicadas a seguir.

Etapa 1: Definindo as questões de pesquisa. Inicialmente, buscou-se pesquisar as questões que dão uma visão geral acerca do estado da arte na área de Forense Digital e Bancos de Dados. As questões de pesquisa e seus propósitos estão descritas no Quadro 1.

Quadro 1. Questões de pesquisa e o propósito de cada uma delas.

Questão	**Propósito**
Quando e onde os estudos foram publicados?	Definindo interesses e tendências ao longo do tempo.
Quais tipos de pesquisa foram feitos?	Classificando em qualitativa, quantitativa ou mista.
Quais conjuntos de dados são considerados?	Definição dos focos de estudo e contextos.
Quais aspectos da forense digital foram enfocados?	Definição das subáreas, temas e tendências mais abordadas.
Que conhecimento foi descoberto?	Identificar métodos, modelos e ferramentas propostos.
Quais problemas foram apontados?	Reconhecendo novos desafios.

Etapa 2: Definindo as strings de pesquisa. Nesta etapa, foram definidas as strings de busca a serem usadas na obtenção dos resultados para responder às perguntas formuladas na primeira etapa. Para encontrar esses dados, foi iniciada uma busca na DBLP (*Digital Bibliography and Library Project*) utilizando o termo de busca “data forense” no intuito de encontrar as principais publicações no campo da computação que envolvessem forense digital e bancos de dados. Em seguida, foram selecionadas as palavras-chave mais relevantes de tais publicações e criamos as seguintes strings de busca: String 1: ”database forensic” OR ”database forensics” OR ”forensic database” OR ”forensic databases”; String 2: ”criminal database” OR ”criminal databases” OR ”database auditing”; String 3: (database OR databases) AND (”forensic access” OR ”forensic analysis” OR ”forensic purpose” OR ”forensic purposes”); String 4: (forensic OR forensics) AND (”database analysis” OR ”database access”); e String 5: (forensic OR forensics) AND (SQL OR NoSQL)

Etapa 3: Definindo o critério de inclusão e o critério de exclusão geral dos dados. Nesta etapa foram definidos os critérios de inclusão e critérios gerais de exclusão para filtrar e excluir publicações irrelevantes. Os critérios foram definidos dessa forma: *i)* Critério de inclusão: O estudo está relacionado com a área de bases de dados que discutem a forense digital; e *ii)* Critério de exclusão: A publicação não possui resumo, foi publicada apenas como resumo, é uma versão antiga de outro estudo considerado, não é um estudo primário ou não disponibiliza o acesso ao estudo completo.

Etapa 4: Etapa de Pesquisa. Para buscar pelas publicações, optamos por realizar as buscas automáticas nas bibliotecas digitais IEEE Xplore, Scopus, Science Direct e Web of Science. Todas as publicações que retornaram da busca nessas bibliotecas foram armazenadas, com exceção daquelas coletadas da biblioteca Scopus, onde apenas as publicações sobre Computing e Engineering foram selecionadas. Ao todo, 5,671 publicações foram coletadas. Os dados sobre as bibliotecas digitais analisadas e suas respectivas quantidades de publicações extraídas estão detalhados no Quadro 2.

Quadro 2. Publicações por biblioteca que retornaram da busca automática.

Biblioteca	Quantidade
IEEE	278
Scopus	3,029
Science Direct	1,665
Web of Science	175

Etapa 5: Definindo critérios de exclusão específicos. A partir da análise dos títulos das publicações encontradas, alguns critérios específicos de exclusão foram criados para restringir as áreas de estudo de interesse. O critério específico de exclusão foi definido da seguinte forma: publicações que não estão inclusas nas áreas da *Computing* ou *Engineering* e publicações que tratam a respeito de várias áreas (multimídia) foram excluídas.

Etapa 6: Selecionando as publicações. Primeiramente, o critério de exclusão foi usado para ler o título e as palavras chaves dos estudos. Posteriormente, foi analisado o resumo das 483 publicações restantes após a aplicação do critério de exclusão. Após a leitura, os artigos que não condizem com o critério de inclusão foram excluídos. Por fim, foram selecionadas 141 publicações únicas.

Etapa 7: Classificação das publicações. Para classificar as publicações, 4 conceitos foram utilizados, são eles: SQL Injection, DBMS, data crawler e forensic knowledge. A partir dos trabalhos selecionados na etapa 6, três voluntários classificaram manualmente as 141 publicações e excluíram aquelas que estavam fora de contexto, restando apenas 91 publicações. Em seguida, para verificar a concordância dessas classificações, aplicamos o coeficiente Kappa de Fleiss Cohen (1960), que chegou a um valor de 0.30, com 95% de credibilidade. Os voluntários, então, discutiram o conteúdo selecionado e chegaram a um consenso que resultou na melhor classificação e na exclusão de mais 21 publicações. Por conseguinte, os resultados dessa revisão sistemática são apresentados para 70 publicações.

RESULTADOS E DISCUSSÕES:

Nesta seção, são descritos os quatro conceitos utilizados para categorizar as publicações em forense digital e bancos de dados. Também são exemplificados trabalhos classificados em cada conceito, conforme mostra a Tabela 1.

Tabela 1. Resumo das publicações encontradas para cada categoria.

Definição	Publicações
SGBD	Man Qi (2014); Wu, K et al. (2014); Choi e Lee (2021); Wagner, J. (2019); Schmitt, S. (2018)
SQL Injection	Thakkar, A., Lohiya, R. (2020); Bauer, D. et al. (2020); Dorai, G., Powell, C. (2020); Kanta, A. et al. (2020); Csaba, B. et al. (2019)
Data Building	Grajeda, C. et al. (2018), Bahjat, A., Jones, J. (2019), Liebler, L. et al. (2019), Freiling, F., Hösch, L. (2018); Awasthi, A. et al. (2018)
Forensic Intelligence	Dimitriadis, A. et al. (2020); Ryser, E. et al. (2020); Sharma, P. et al. (2020); Schneider, J. et al. (2020); Zhang, X. et al. (2020)

SGBD (Sistema de Gerenciamento de Banco de Dados) é um conjunto de softwares que ajudam a gerenciar uma base de dados. Ele tem como função armazenar, modificar e realizar a extração de arquivos em banco de dados, além de tornar seguro as operações que ocorrem em sua interface. Ele possui o próprio sistema de segurança. Além de tornar o compartilhamento de dados muito mais simples, possuindo também o controle de redundância, que serve para evitar que os dados se repitam.

SQL Injection é um tipo de ataque da web usado para infectar um banco de dados, que explora vulnerabilidades existentes em aplicativos da web ou mesmo em procedimentos armazenados de servidor de banco de dados em nível de back-end. Em outras palavras, a injeção de SQL permite que

os invasores injetem consultas SQL não seguras para alterar o objetivo anterior da consulta. Os invasores podem obter acesso não autorizado a um banco de dados, ler dados protegidos, corromper o banco de dados ou até mesmo conceder acesso a outros usuários não autorizados. Encontramos trabalhos que identificam como esses ataques podem impactar a perícia digital.

Data Building é um processo que envolve a coleta de informações, agrupando as mesmas, diversificando e segmentando os dados. O levantamento de dados e a posterior análise e esquematização dos mesmos ditam o rumo do trabalho e auxiliam os pesquisadores na construção dos resultados almejados. Dentre as diversas formas de se obter dados para as pesquisas, destacam-se o *Data Scraping* e o *Data Crawling*. *Data Scraping* pode ser entendido como um processo de extração e esquematização de dados retirados majoritariamente da web, como os sites informativos, mas também de diversos outros tipos de aplicações e dispositivos de armazenamento que contenham informações dispostas de maneira não estruturada, tornando esses dados legíveis, maleáveis e úteis ao usuário. Analogamente, o *Data Crawling* também está relacionado à extração e segmentação de dados extraídos de fontes de dados. No entanto, apesar de possuírem semelhanças objetivacionais, a diferença entre os dois se encontra na dimensão dos dados analisados e na especificidade da extração, pois o primeiro consegue filtrar os dados de maneira muito mais específica, uma vez que os mecanismos de filtragem utilizados pelos bots que realizam *Data Crawling* são frequentemente destinados à análise de bases de dados extensas.

Forensic Intelligence é a área da ciência forense interessada e preocupada em encontrar a relação entre pessoas, lugares e coisas envolvidas em atividades criminosas. É uma importante ferramenta para auxiliar em investigar e julgar casos civis e criminais. Neste trabalho, o foco está na forense digital, uma ciência cujo objetivo é identificar, preservar, recuperar, analisar e apresentar dados digitais durante uma investigação digital. Embora os termos forense computacional, forense digital e forense cibernética sejam usados indistintamente, eles são diferentes: a forense computacional é principalmente sobre a investigação de crimes em que os computadores estão relacionados, enquanto a forense cibernética e digital se refere principalmente a dados digitais de diferentes dispositivos digitais. As publicações aqui são classificadas como inteligência forense quando seu objetivo é duplo: analisar o conteúdo do banco de dados para investigar os incidentes do banco de dados e construir uma linha do tempo das atividades criminosas; e adquirir um produto a partir do processamento lógico de dados de casos forenses.

CONCLUSÕES:

As publicações retornadas sobre forense digital têm sido cada vez maiores, além da realização de perícia forense nos mais variados meios computacionais e digitais. Com a pesquisa, foi possível estudar e descrever as diferentes funções que a forense digital tem ocupado. O volume de pesquisas

científicas referentes à área de forense digital tende a aumentar cada vez mais, levando em conta a evolução da computação e a importância da investigação de crimes digitais, que vem sendo cada vez mais frequentes. Uma das maiores limitações no andamento da pesquisa foi referente a classificação das categorias, dado que não há uma organização dos artigos científicos nas bibliotecas acadêmicas. Não obstante, os resultados retornados pelas strings criadas a partir das classificações, necessitavam de uma seleção mais detalhada. Houve também certa dificuldade de acessar algumas pesquisas, uma vez que as mesmas necessitavam de permissão a domínios específicos. O desenvolvimento de um banco de dados que organize as publicações, de modo a facilitar a pesquisa, é um dos trabalhos futuros para este projeto.

REFERÊNCIAS BIBLIOGRÁFICAS:

Awasthi, A., Read, O. L. H., Xynos, K., Sutherland, I. (2018). Welcome pwn: Almond smart home hub forensics.

Bahjat, A., Jones, J. (2019).Deleted file fragment dating by analysis of allocated neighbors.

Bauer, D. et al. (2020). Building and Operating a Large-Scale Enterprise Data Analytics Platform.

Choi, H., Lee, S. (2021). Forensic Recovery of SQL Server Database: Practical Approach.

Cruz, Diego; RODRIGUES, Juliana. CRIMES CIBERNÉTICOS E A FALSA SENSAÇÃO DE IMPUNIDADE. REVISTA CIENTÍFICA ELETRÔNICA DO CURSO DE DIREITO, [s. l.], ano 2018, Disponível em: http://faef.revista.inf.br/imagens_arquivos/arquivos_destaque/iegWxiOtVJB1t5C_2019-2-28-16-36-0.pdf. ed. 13ª Edição, p. 0-18, Jan 2018.

Csaba, B.,Tamás, H., Horváth, A., Oláh, A., Reguly, Z. I. (2019). PPCU Sam: Open-source face recognition framework.

Dimitriadis, A., Ivezic, N., Kulvatunyou, B., and Mavridis, I. (2020). D4i - digital forensics framework for reviewing and investigating cyber attacks. Array, 5:100015.

Dorai, G., Powell, C. (2020). VIDE - Vault App Identification and Extraction System for iOS Devices.

Freiling, F., Hösch, L. (2018). Controlled experiments in digital evidence tampering.

Grajeda, C. et al. (2018). Experience constructing the Artifact Genome Project (AGP): Managing the domain's knowledge one artifact at a time.

Gonçalves, M. et al. Perícia Forense Computacional: Metodologias, Técnicas e Ferramentas. Disponível em: http://faef.revista.inf.br/imagens_arquivos/arquivos_destaque/iegWxiOtVJB1t5C_2019-2-28-16-36-0.pdf. Mato Grosso: Revista Científica Eletrônica de Ciências Sociais Aplicadas da EDUVALE, nov, 2012.

Jaques, Gabriel et al. UMA ANÁLISE SOBRE MALWARES E ESTRATÉGIAS DE PREVENÇÃO. Disponível em: Disponível em: http://faef.revista.inf.br/imagens_arquivos/arquivos_destaque/iegWxiOtVJB1t5C_2019-2-28-16-36-

0.pdf. XVIII Seminário Internacional de Educação no MERCOSUL, Cruz Alta, RS, ano 2018, ed. 18ª Edição, p. 1-4, 10 maio 2018.

Kanta, A., Coisel, I., Scanlon, M. (2020). A survey exploring open source Intelligence for smarter password cracking.

Kitchenham, B., & Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering.

Liebler, L., Schmitt, P., Baier, H., Breitinger, F. (2019). On efficiency of artifact lookup strategies in digital forensics.

Man Qi (2014). Digital Forensics and NoSQL Databases.

Ryser, E., Spichiger, H., and Casey, E. (2020). Structured decision making in investigations involving digital and multimedia evidence. Forensic Science International: Digital Investigation, 34:301015.

Sharma, P., Arora, D., and Sakthivel, T. (2020). Enhanced forensic process for improving mobile cloud traceability in cloud-based mobile applications. Procedia Computer Science, 167:907–917. International Conference on Computational Intelligence and Data Science.

Schmitt, S. (2018). Introducing Anti-Forensics to SQLite Corpora and Tool Testing.

Schneider, J., Wolf, J., and Freiling, F. (2020b). Tampering with digital evidence is hard: The case of main memory images. Forensic Science International: Digital Investigation, 32:300924.

Thakkar, A. and Lohiya, R. (2020). A review of the advancement in intrusion detection datasets. Procedia Computer Science, 167:636–645. International Conference on Computational Intelligence and Data Science

Wagner, James et al. (2019). DB3F&DF-Toolkit: The Database Forensic File Format and theDatabase Forensic Toolkit. .

Wu, K.; Hua, L.; Wang, X.; Ding, X. (2014). The design and implementation of database audit system framework.

Zhang, X., Upton, O., Beebe, N. L., and Choo, K.-K. R. (2020). Iot botnet forensics: A comprehensive digital forensic case study on mirai botnet servers. Forensic Science International: Digital Investigation, 32:300926.

Participação em Congressos, publicações e/ou pedidos de proteção intelectual:

SEUFITELLI, D. B. ; MATA, W. R. R. ; SOUZA, R. C. BRANDÃO M. A.; MORO, M. M. Characterization and Analysis of Open Brazilian Judiciary Data. In: Webmedia 2020 - 26º Simpósio Brasileiro de Sistemas Multimídia e Web, 2020, São Luís. Anais do 26º Simpósio Brasileiro de Sistemas Multimídia e Web. Porto Alegre: SBC, 2020.

MATA, W. R. R.; SOUZA, R. C.; FERNANDES, A. C. A.; BRANDÃO M. A. Tojudge: Forense Digital E Dados Abertos Do Poder Judiciário Brasileiro. SNCT/IFMG 2020.