Forense Digital e Bancos de Dados: um Survey

Danilo B. Seufitelli[1], Ana Flávia C. Moura[1], Ayane C. A. Fernandes[2], Kayque M. Siqueira[2], Michele A. Brandão[1,2], Mirella M. Moro[1]

[1]Universidade Federal de Minas Gerais – Belo Horizonte – MG

[2]Instituto Federal de Minas Gerais – Ribeirão das Neves – MG

{daniloboechat, anaciriaco, michele.brandao, mirella}@dcc.ufmg.br,
{ayanecristinamb, kayque.siq}@gmail.com

Abstract. *This work summarizes a systematic literature review with a simple a classification for publications at the intersection between Digital Forensics and Databases. This research opens the doors for the communication between databases and an area that has several interesting and concrete challenges.*

Resumo. *Este artigo resume uma revisão sistemática da literatura com classificação simples para publicações na interseção entre Forense Digital e Bancos de Dados. Esta pesquisa abre as portas para comunicação entre Bancos de Dados e uma área com vários desafios interessantes e concretos.*

1. Introdução

Este artigo resume uma pesquisa original sobre Bancos de Dados (BD) e suas intersecções com Forense Digital. A Forense Digital é uma ciência crucial no cenário de cibercrimes, pois auxilia na reconstrução dos mesmos durante investigação e no desenvolvimento de medidas para prevenção de suas ocorrências. Especificamente, essa ciência atua na busca, análise, identificação e categorização de dados que possam ser considerados evidências de crimes [Al-Dhaqm et al. 2020a, Qi et al. 2014]. A Forense Digital tem sido utilizada na prevenção e detecção de ataques de *SQL Injection* [Xie et al. 2019], por exemplo.

Existem muitas pesquisas e tecnologias na área de Forense Digital,[1] para tratar esse grande volume de publicações, é necessário aplicar uma metodologia de Revisão Sistemática da Literatura (RSL) [Kitchenham and Charters 2007] a fim de identificar as mais pertinentes no contexto de BD (Seção 2). Aqui, são apresentadas as publicações resultantes da RSL e como os dados são utilizados na Forense Digital (Seção 3), além de uma discussão sobre oportunidades de pesquisa na sua interseção com BD (Seção 4).

2. Metodologia

Neste trabalho, a metodologia utilizada se baseia na execução de sete etapas adaptadas do protocolo de Kitchenham e Charters (2007), conforme resumido a seguir.

Etapa 1: Definir questões de pesquisa. Elaboramos cinco questões exploratórias de pesquisa (Tabela 1) para obter uma visão geral da literatura de Forense Digital no contexto de BD e auxiliar na descrição e classificação dos estudos selecionados para a RSL.

[1]Embora os termos forense computacional, forense digital e forense cibernética sejam usados indistintamente, eles são diferentes: forense computacional foca na investigação de crimes em que os computadores estão presentes, e a forense cibernética e digital se referem a dados de diferentes dispositivos digitais.

Tabela 1. Questões de pesquisa e strings de busca.

Questões de Pesquisa
Qual a interseção entre bancos de dados e forense digital?
Quais tipos de pesquisa mais comuns realizadas na forense digital: qualitativa, quantitativa ou mista?
Quais conjuntos de dados são considerados em estudos da área de forense digital?
Quais subáreas podem ser identificadas na interseção entre forense digital e bancos de dados?
Quais desafios e oportunidades em trabalhos na interseção entre BD e forense digital?
Strings de Busca
"database forensic" OR "database forensics" OR "forensic database" OR "forensic databases"
"criminal database" OR "criminal databases" OR "database auditing"
(database OR databases) AND ("forensic access" OR "forensic analysis" OR "forensic purpose" OR "forensic purposes")
(forensic OR forensics) AND ("database analysis" OR "database access")
(forensic OR forensics) AND (SQL OR NoSQL)

Etapa 2: Definir strings de pesquisa. A partir dessas questões, as strings de busca foram definidas. A parte inferior da Tabela 1 informa o conjunto final de strings.

Etapa 3: Definir critério de inclusão e de exclusão geral dos dados. Esses critérios auxiliam na definição do conjunto de publicações: para inclusão, o conteúdo está relacionado com a área de BD e discute forense digital; e para exclusão, a publicação não possui resumo, é apenas um resumo, é uma versão antiga de outro estudo considerado, não é um estudo primário ou não disponibiliza o acesso ao estudo completo, ou é duplicada.

Etapa 4: Pesquisar. A busca por publicações considerou as bibliotecas digitais: IEEE Xplore, Scopus, Science Direct e Web of Science. Foram 5,671 artigos coletados: IEEE 278 artigos, Scopus 3.029, Science Direct 1.665, e Web of Science 175.

Etapa 5: Definir critérios de exclusão específicos. A partir dos títulos, criaram-se critérios específicos de exclusão: publicações fora de *Computing* e de *Engineering*, e as que tratam de outras áreas – e.g., trabalhos de biologia como *genetic forense* e biomedicina. Após aplicar tais critérios em título e palavras-chave, restaram 483 publicações.

Etapa 6: Selecionar publicações e identificar temas comuns. Pela leitura do resumo, descartaram-se artigos fora dos critérios de inclusão, restando 141 trabalhos. Foi realizada leitura dinâmica desses e identificados dois temas predominantes.

Etapa 7: Classificar publicações. Os temas identificados no passo anterior foram conceituados para elaboração de uma taxonomia. Então, três voluntários[2] classificaram manualmente as 141 publicações. Em paralelo, uma reanálise de exclusão foi aplicada e publicações fora do escopo foram excluídas. Com a nova filtragem, apenas 91 publicações foram consideradas. Os voluntários, então, discutiram o conteúdo dos trabalhos e chegaram ao consenso que resultou em 70 publicações, das quais 29 são o foco deste resumo.[3]

3. Análises e Discussões das Publicações

Este artigo resume as publicações melhor relacionadas à area de BD, sendo classificadas em: Construção de Dados e Sistema de Gerenciamento de Bancos de Dados.

Construção de Dados. Das **16** publicações classificadas como Construção de Dados, **oito** construíram seu BD a partir de dados retirados de aplicativos e dispositivos externos, conforme a Tabela 2. Em [Awasthi et al. 2018, Atwal et al. 2019,

[2]Professor/Doutor, Doutorando e Graduando em Ciência da Computação com experiência em BD.

[3]Planilha com as publicações selecionadas: `https://bit.ly/papersForense`

Tabela 2. Publicações classificadas como Construção de Dados.

Referência Palavras-chave	Fonte de Dados
⊕ [Khobragade and Malik 2014] Geração/mineração de dados	– Dados coletados de logs de navegadores e fluxos de dados em redes e computadores.
⊙ [Li et al. 2014] Android Recov. Methods	– Base de dados de SMS de Samsung Galaxy S3, considerando a partição de dados no dispositivo em três períodos
⊙ [Satrya et al. 2016] Android Telegram	– Base de dados extraída de artefatos de mensagens de texto e áudio do aplicativo Telegram de três dispositivos móveis Android
⊙ [Awasthi et al. 2018] Smart home hub	– Dados extraídos de hub doméstico (*Securifi Almond*), que permitiu a análise forense de artefatos relativos à interação do usuário no hub
⊘ [Freiling and Hösch 2018] Adulteração de evidências	– Dados em uma imagem de disco foram manipulados por estudantes de pós-graduação para estudo de evidências digitais adulteradas
⊕ [Atwal et al. 2019] Spotlight, Apple desktop	– Base com metadados do *Spotlight* (ferramenta de busca, Apple) para pesquisar persistência de dados excluídos nos seus metadados
⊘ [van Zandwijk and Boztas 2019] iPhone Health App	– Investiga o uso de dados extraídos do aplicativo Health do Iphone, que armazena dados sobre exercícios físicos realizados pelo usuário
⊕ [Servida and Casey 2019] IoT, digital traces	– Dados provenientes de aplicativos de celular (e.g., hubs Nest e Wink) e de dispositivos IoT (e.g., dados da nuvem do QBee Câmera e o Swisscom Home App); foram desenvolvidos plugins para extração de dados

⊙ Extração de dados de dispositivos. ⊕ Recuperação de dados. ⊘ Evidência digital.

van Zandwijk and Boztas 2019], o foco é viabilizar o uso de dados extraídos de dispositivos como ferramenta para análise forense. Por exemplo, van Zandwijk e Boztas (2019) apresentam experimentos em três dispositivos iOS sobre o aplicativo *Health*, que armazena dados sobre a quantidade de passos e tempo em movimento da pessoa.

Sobre outras ferramentas específicas, Satrya et al. (2016) usam duas para retirar as mensagens de texto e áudio do aplicativo Telegram de três dispositivos móveis Android. Ao final, os dados remanescentes no armazenamento do Telegram podem ser extraídos e usados como evidência digital de crimes cibernéticos. Servida e Casey (2019) apresentam como vestígios de dados de dispositivos IoT (*Internet of Things*) podem ser úteis à forense. Dispositivos estudados incluem *QBee Multi-Sensor Camera*, *Cube One & Accessories*, *Arlo Pro* e o *Nest Protect*. Dados desses aparelhos foram extraídos com múltiplos plugins desenvolvidos pelos autores e de códigos abertos (e.g., *Autopsy Framework*).

Completando, Li et al. (2014) apresentam um método de recuperação de registros operacionais de arquivos, e Khobragade e Malik (2014) propõem uma metodologia para coletar e analisar dados de sistemas cibernéticos e logs de navegadores. Igualmente, Liu et al. (2016) apresentam um método de recuperação e análise de dados sobre os arquivos apagados de BD do SQLite3 (software de BD utilizado por celulares). Nesse estudo, a localização e o tamanho dos dados no campo pesquisado são estimados a partir da análise dos formatos SQLite3. Finalmente, Freiling e Hosch (2018) descrevem experimentos realizados para diferenciar evidências adulteradas e não adulteradas.

A análise dessas publicações revela que o foco é obter dados a partir de dispositivos móveis e aplicativos. Observa-se que Android é o sistema mais usado nesses estudos.

Sistema de Gerenciamento de Bancos de Dados. Das **13** classificadas como SGBD, as **cinco** mais relevantes são resumidas na Tabela 3 com respectiva descrição do SGBD utilizado. Hommes et al. (2013) usam um BD NoSQL para armazenar registros de parâmetros de rede a fim de viabilizar o gerenciamento de falhas online. Mais amplamente, Qi (2014) analisa os quatro tipos principais de BD NoSQL, mas foca na análise de desempenho do MongoDB e Riak. Similarmente, Qi et al. (2014) analisam duas técnicas para melhorar a

Tabela 3. Publicações classificadas como SGBD.

Referência Palavras-chave	Fonte de Dados
⊕ [Hommes et al. 2013] Source Code, Debug	– Propõe um framework para segurança de dados em rede e os registros são feitos em um banco de dados NoSQL
⊙ [Qi 2014] NoSQL Databases	– Analisa quatro SGBDs NoSQL, com foco na análise de desempenho do MongoDB e Riak
⊙ [Qi et al. 2014] Big Data	– Compara MongoDB, Riak e MySQL quanto ao gerenciamento de dados
⊙ [Khanji et al. 2015] Database Auditing	– Utiliza os SGBDs Oracle e MySQL para testar o desempenho de um framework de busca proposto
⊘ [Choi et al. 2021] Forensic Recovery	– Introduz um método de recuperação de dados apagados para ser utilizado no MySQL

⊙ Análise de desempenho. ⊕ Regras de segurança. ⊘ Recuperação de dados.

escalabilidade no gerenciamento de dados, comparando MongoDB, Riak e MySQL.

Mudando para BD relacional, Khanji et al. (2015) focam no MySQL e no Oracle, e propõem um framework com funcionalidades de auditoria para auxiliar na realização de análises em bancos de dados forenses. MySQL também é foco de Choi et al. (2021), que verificam o desempenho de um método proposto para recuperar arquivos apagados. Esse teste considera conjuntos de dados criados pelos autores com características específicas.

Note que Khanji et al. (2015) trazem um conceito importante de **BD Forense**, que pode ser uma subárea da forense digital com pouco foco na literatura. Esses bancos de dados forenses são logicamente idênticos aos comuns, mas diferem quanto à estrutura física do arquivo, mecanismos de segurança e concorrência, otimização de consultas, entre outros. Em particular, a área de bancos de dados forense requer o desenvolvimento de ferramentas de análise forense que possam ser utilizadas em diferentes SGBDs.

4. Discussão e Oportunidades

Esta seção sumariza os resultados da RSL com respostas para cada questão de pesquisa.

Qual a interseção entre bancos de dados e forense digital? A principal interseção é a necessidade que a Forense Digital tem em relação a seus dados, os quais geralmente precisam ser armazenados durante a investigação e para futuros estudos. Essa necessidade é geralmente adequada frente ao objetivo da pesquisa. Em resumo, os dados são utilizados para a própria investigação forense [Ming and LiZhong 2009, Al-Dhaqm et al. 2020a, Al-Dhaqm et al. 2020b], e para testar o desempenho de uma abordagem proposta (metodologia, framework ou linguagem) [Choi et al. 2021, Khanji et al. 2015]. Também é possível a proposição de um conjunto de dados [Awasthi et al. 2018, Atwal et al. 2019, van Zandwijk and Boztas 2019]. Conclui-se que não existe uma escolha em relação ao modelo de dados utilizado, pois trabalhos como [Hommes et al. 2013, Qi 2014, Qi et al. 2014] usam bancos NoSQL, e [Khanji et al. 2015, Choi et al. 2021] o modelo relacional, sendo MongoDB e MySQL os SGBDs que mais são mencionados.

Quais tipos de pesquisa mais comuns realizadas na forense digital: qualitativa, quantitativa ou mista? A maioria das pesquisas nos trabalhos selecionados são quantitativas, ou seja, utilizam diferentes estratégias estatísticas para validar determinada hipótese. Entre as poucas exceções estão [Freiling and Hösch 2018, Henseler and van Loenhout 2018, Servida and Casey 2019], que utilizam uma abordagem qualitativa.

Quais conjuntos de dados são considerados em estudos da área de forense digital? Os trabalhos selecionados propõem diversos conjuntos de dados novos e também utilizam conjuntos de dados já existentes. Com pesquisas bastante diversas, há trabalhos que utilizam dados de fluxos de pacotes em redes de computadores [Sikos 2020], de registros de automóveis roubados [Chen 2008], e de logs de navegadores [Khobragade and Malik 2014, Salunkhe et al. 2016].

Quais subáreas podem ser identificadas na interseção entre forense digital e bancos de dados? Foram identificadas duas classes de pesquisas, cada uma com três categorias: *construção de dados* – extração de dados de dispositivos, recuperação de dados e evidências digitais; e e *SGBD* – análise de desempenho, regras de segurança e recuperação de dados. Apesar das classes terem a categoria de *recuperação de dados*, o foco em cada uma delas é diferente: em construção de dados, a recuperação foca na construção dos dados e estudos associados a eles (e.g., análise da persistência dos dados quando excluídos e desenvolvimento de plugins para extração de dados); e em SGBD, a recuperação de dados está associada aos sistemas em si (e.g., recuperação de dados no MySQL).

Quais desafios e oportunidades em trabalhos na interseção entre BD e Forense Digital? Um desafio é a mudança de tecnologias, pois cada vez surgem mais formatos de dados, de sistemas, de linguagens de programação, de dispositivos físicos, entre outros. Em relação a BD, um desafio é manter a compatibilidade entre sistemas. De outro modo, novas tecnologias também demandam a formação de mais especialistas e criação de novas abordagens para a investigação forense. Por exemplo, muitos trabalhos focam na extração de dados em dispositivos com sistemas específicos, como android, IOS, câmeras, videogames, e outros. Finalmente, conforme Kanji et al. (2015), um desafio relevante é buscar o equilíbrio entre o desempenho de um SGBD e os recursos de auditoria para investigação forense. Especificamente, é preciso tornar o banco de dados ajustável para tais análises.

5. Conclusão

O número de publicações sobre forense digital tem aumentado, e a realização de perícia forense se ampliado em vários contextos computacionais e digitais. Como área, Bancos de Dados tem muito a colaborar com a realização de investigações forense, por exemplo, por meio da busca por maior velocidade de processamento. A RSL também revelou que técnicas mais avançadas de organização e processamento de dados ainda não foram exploradas mais amplamente na investigação criminal. O próximo passo desta pesquisa é ampliar a cobertura de publicações consideradas para áreas correlatas a Bancos de Dados, como Mineração de Dados e Aprendizado de Máquina, entre outras.

Agradecimentos. Este trabalho foi parcialmente financiado pela CAPES e por recursos do Edital de pesquisa 087/2019 do IFMG.

Referências

Al-Dhaqm, A. et al. (2020a). Categorization and organization of database forensic investigation processes. *IEEE Access*, 8:112846–112858.

Al-Dhaqm, A. et al. (2020b). Database forensic investigation process models: A review. *IEEE Access*, 8:48477–48490.

Atwal, T. S. et al. (2019). Shining a light on spotlight: Leveraging apple's desktop search utility to recover deleted file metadata on macos. *Digital Investigation*, 28:S105–S115.

Awasthi, A. et al. (2018). Welcome pwn: Almond smart home hub forensics. *Digital Investigation*, 26:S38–S46.

Chen, P. S. (2008). *Discovering Investigation Clues through Mining Criminal Databases*, pages 173–198. Springer Berlin Heidelberg, Berlin, Heidelberg.

Choi, H., Lee, S., and Jeong, D. (2021). Forensic recovery of SQL server database: Practical approach. *IEEE Access*, 9:14564–14575.

Freiling, F. and Hösch, L. (2018). Controlled experiments in digital evidence tampering. *Digital Investigation*, 24:S83–S92.

Henseler, H. and van Loenhout, S. (2018). Educating judges, prosecutors and lawyers in the use of digital forensic experts. *Digital Investigation*, 24:S76–S82.

Hommes, S. et al. (2013). Automated source code extension for debugging of openflow based networks. In *CNSM*, pages 105–108.

Khanji, S. I. R., Khattak, A. M., and Hacid, H. (2015). Database auditing and forensics: Exploration and evaluation. In *AICCSA*, pages 1–6.

Khobragade, P. K. and Malik, L. G. (2014). Data generation and analysis for digital forensic application using data mining. In *CSNT*, pages 458–462.

Kitchenham, B. and Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE-2007-01, Keele University and Durham University Joint Report.

Li, Q., Hu, X., and Wu, H. (2014). Database management strategy and recovery methods of android. In *ICSESS*, pages 727–730.

Liu, X., Fu, X., and Sun, G. (2016). Recovery of deleted record for SQLite3 database. In *IHMSC*, pages 183–187.

Ming, H. and LiZhong, S. (2009). A new system design of network invasion forensics. In *ICCEE*, pages 596–599.

Qi, M. (2014). Digital forensics and NoSQL databases. In *FSKD*, pages 734–739.

Qi, M. et al. (2014). Big data management in digital forensics. In *CSE*, pages 238–243.

Salunkhe, P., Bharne, S., and Padiya, P. (2016). Data analysis of file forensic investigation. In *SCOPES*, pages 372–375.

Satrya, G. B., Daely, P. T., and Nugroho, M. A. (2016). Digital forensic analysis of telegram messenger on android devices. In *ICTS*, pages 1–7.

Servida, F. and Casey, E. (2019). Iot forensic challenges and opportunities for digital traces. *Digital Investigation*, 28:S22–S29.

Sikos, L. F. (2020). Packet analysis for network forensics: A comprehensive survey. *FSI: Digital Investigation*, 32:200892.

van Zandwijk, J. P. and Boztas, A. (2019). The iphone health app from a forensic perspective: can steps and distances registered during walking and running be used as digital evidence? *Digital Investigation*, 28:S126–S133.

Xie, X. et al. (2019). SQL injection detection for web applications based on elastic-pooling cnn. *IEEE Access*, 7:151475–151481.