UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Um *framework* baseado em blockchain para preservar a privacidade no compartilhamento de dados de saúde

Erikson Júlio de Aguiar

Dissertação de Mestrado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Erikson Júlio de Aguiar

Um *framework* baseado em blockchain para preservar a
privacidade no compartilhamento de dados de saúde

Dissertação apresentada ao Instituto de Ciências
Matemáticas e de Computação – ICMC-USP,
como parte dos requisitos para obtenção do título
de Mestre em Ciências – Ciências de Computação e
Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e
Matemática Computacional

Orientador: Prof. Dr. Jó Ueyama

USP – São Carlos
Fevereiro de 2021

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

A282f   Aguiar, Erikson Júlio de
            Um framework baseado em blockchain para
preservar a privacidade no compartilhamento de
dados de saúde / Erikson Júlio de Aguiar;
orientador Jó Ueyama. -- São Carlos, 2021.
            124 p.

            Dissertação (Mestrado - Programa de Pós-Graduação
em Ciências de Computação e Matemática
Computacional) -- Instituto de Ciências Matemáticas
e de Computação, Universidade de São Paulo, 2021.

            1. Blockchain. 2. Privacidade. 3. Ataques de Re-
identificação. 4. Dados de Saúde. 5. Hyperledger
Fabric. I. Ueyama, Jó , orient. II. Título.

Erikson Júlio de Aguiar

**A blockchain-based framework towards privacy preserving in
health data sharing**

Dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP – in accordance with the requirements of the Computer and Mathematical Sciences Graduate Program, for the degree of Master in Science. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Jó Ueyama

USP – São Carlos
February 2021

Aos meus pais, Silvana e Edson.

AGRADECIMENTOS

Agradeço primeiramente aos meus pais, que me apoiaram durante esses últimos dois anos de estudos e me ajudaram a superar desafios.

Agradeço ao meu orientador, Professor Dr. Jó Ueyama pela oportunidade em desenvolver esta pesquisa, por toda atenção, paciência e colaboração.

Ao **Instituto de Ciências Matemáticas e de Computação (ICMC)** da Universidade de São Paulo (USP), por toda a infraestrutura disponibilizada no desenvolvimento desta pesquisa.

Ao **Laboratório de Sistemas Web e Multimídia Interativos (Intermídia)**, pelo acolhimento e espaço oferecido, que contribuíram para o progresso deste trabalho. Além disso, um agradecimento especial aos meus amigos do Intermídia, por todas as sugestões, apoio, conselhos e companheirismo, durante esses dois anos de pesquisa.

Aos meus amigos, Thiago Costa, Alef Vinicius, Felipe Giuntini, Heitor Freitas, José Torres Neto, Kauê Moraes e Lucas Brito, do grupo *Wines* que contribuíram de forma direta ou indireta com este trabalho. Além disso, agradeço a todos os meus amigos do laboratório Intermidia que contribuíram de alguma forma com este trabalho. Um agradecimento especial ao Prof. Dr. Bruno Faiçal, que me ajudou muito no início do mestrado e pelas contribuições nos artigos. Também, gostaria de agradecer ao Prof. Dr. Alyson de Jesus, pela ajuda durante o desenvolvimento desta pesquisa, bem como, pelas contribuições técnicas e científicas para este trabalho.

À **Coordenação de Aperfeiçoamento Pessoal de Nível Superior (CAPES) - código de financiamento 001**, pelo apoio financeiro, na modalidade bolsa de mestrado, durante o início deste trabalho.

À **Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP)** pelo apoio financeiro, na modalidade bolsa regular de mestrado - **processo nº 2018/18187-3**.

*“Sonhos determinam o que você quer.
Ação determina o que você conquista.”
(Aldo Novak)*

RESUMO

AGUIAR, E. J. **Um *framework* baseado em blockchain para preservar a privacidade no compartilhamento de dados de saúde.** 2021. 124 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2021.

O compartilhamento de dados de saúde para pesquisa contribui para gerar novos conhecimentos. Na pandemia do novo coronavírus, sua importância é ressaltada, pois, pesquisadores de todo o mundo trocam informações para contribuir na busca de uma vacina efetiva contra o vírus. A blockchain pode auxiliar no compartilhamento de dados de saúde, reduzindo riscos a fraudes, a partir das características como imutabilidade, não repúdio e confiabilidade. No entanto, as organizações de saúde ainda estão suscetíveis a vazamento, falhas e ataques cibernéticos. Entre os ataques sofridos, um deles é o ataque de re-identificação que tem o propósito de associar registros públicos dos usuários, de modo a encontrar padrões para descobrir a identidade deles. Sistemas de saúde baseados em blockchain também estão passíveis a ataques de re-identificação. Portanto, esta pesquisa tem como suposições: (i) a adição de técnicas de anonimização ao blockchain podem mitigar riscos à privacidade dos usuários; (ii) a utilização de um *token* (valor *hash* que é utilizado como chave para auditoria no blockchain) pode auxiliar no rastreamento dos ativos compartilhados. Esta pesquisa tem o objetivo de desenvolver um *framework* blockchain com o *Hyperledger Fabric* para preservar a privacidade no compartilhamento de dados de saúde. Em vista disso, foram empregados os métodos de anonimização K-Anonimato e Privacidade Diferencial para compô-lo e mitigar problemas à privacidade. Além disso, este trabalho elaborou um estudo piloto para o compartilhamento de imagens médicas frente a ataques de re-identificação. Os resultados revelam um aumento nas despesas para a rede e hardware dos sistemas de saúde, quando utilizam o *framework* proposto, devido ao custo gerado pelo blockchain e pelos modelos de privacidade. No entanto, as despesas são justificáveis, uma vez que, o *framework* acresce aos sistemas de saúde em propriedades como não repúdio, confiabilidade e privacidade. As análises da privacidade, que foram estatísticas e para as funções de probabilidade, revelam que o comportamento de ambos os métodos apresentam baixa probabilidade da identificação de registros similares, expressa pela métrica entropia. Todo o código relativo ao *framework* pode ser encontrado em: <<https://github.com/eriksonJAguiar/bcpshield-framework>>.

Palavras-chave: Blockchain, Privacidade, Ataques de Re-identificação, Dados de Saúde, Hyperledger Fabric.

ABSTRACT

AGUIAR, E. J. **A blockchain-based framework towards privacy preserving in health data sharing**. 2021. 124 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2021.

Sharing health data for research is a crucial practice to generate knowledge. In the Coronavirus pandemic, it highlights data sharing importance because of researchers around the world to exchange information to discover an efficient vaccine against the virus. A technology workable for sharing health data is the blockchain that helps to avoid fraud and can bring benefits such as immutability, non-repudiation, and reliability. At this point, healthcare organizations suffering from data leakage, failure, and cyberattacks, one of them is the re-identification attack. The re-identification attack aims to associate public records of users with finding patterns to discover their identity. Our research considers the following hypothesis: (i) techniques to mitigate privacy issues on blockchain-based systems for sharing health data; (ii) we assume that a tracking token (the hash value which key for auditing on the blockchain) into image metadata can assist to track shared images. The method used to achieve our aim comprises developing a permissioned blockchain framework with the Hyperledger Fabric for preserve privacy in shared health data. Besides, we use the anonymization techniques K-Anonymity and Differential Privacy to aggregate in the framework and mitigate privacy risks. For tests, we adopt a pilot study for sharing medical imaging in front of re-identification attacks. Our results revealed the hardware and network performance for the framework has a higher overhead. However, the results are justifiable because they bring benefits, such as non-repudiation, reliability, and privacy. We apply statistics metrics and probability function to analyze behavior of privacy variables against re-identification attacks on blockchain. Our finds related to privacy reveals that both methods have low probability of identify similar records through the entropy. The source code of the framework can be found at: <https://github.com/eriksonJAguiar/bcpshield-framework>.

Keywords: Blockchain, Privacy, Re-identification Attacks, Healthcare Data, Hyperledger Fabric.

LISTA DE ILUSTRAÇÕES

Figura 1 – Estrutura de dados do blockchain. Adaptado de (DINH <i>et al.</i> , 2018).	34
Figura 2 – Visão geral do <i>framework</i> Hyperledger Fabric. Adaptado de (ANDROULAKI <i>et al.</i> , 2018).	39
Figura 3 – Fluxograma para a escolha do tipo de rede blockchain e processo de consenso. Adaptado de Wust e Gervais (2018); Lo <i>et al.</i> (2017).	55
Figura 4 – Visão geral do <i>framework</i> BCPSHield, que é composto por 11 fluxos para compartilhar e recuperar dados. Os fluxos são: (1) aplicação arbitrária do usuário; (2) interface de comunicação com os demais componentes; (3) canal de envio de dados; (4) observador local; (5) base de dados local; (6) componente de privacidade dos dados; (7) inserir <i>token</i> de rastreamento nos dados compartilhados; (8) rede <i>Off-chain</i> ; (9) rede blockchain; (10) canal de envio dos dados compartilhados; (11) canal de troca de informações compartilhadas com a interface.	72
Figura 5 – Arquitetura do Hyperledger Fabric para a rede blockchain desta proposta. Baseado em Androulaki <i>et al.</i> (2018).	75
Figura 6 – Arquitetura de referência das camadas da blockchain. Adaptado de Yuan e Wang (2018). As siglas apresentadas nas figuras são: K-Anonimato (KA), Criptografia Baseada em Atributos (CBA), Privacidade Diferencial (PD), <i>Mixing</i> (MX), <i>JavaScript Object Notation</i> (JSON), <i>Hyperledger Fabric</i> (HLF), <i>Certificate Authority</i> (CA), <i>Distributed Hash Table</i> (DHT), e <i>InterPlanetary File System</i> (IPFS).	77
Figura 7 – Cenário de estudo utilizado como base para as simulações. É composto pelos médicos, que aguardam os exames dos pacientes; paciente, que autoriza o compartilhamento dos dados pelos provedores de saúde; e pesquisador, que solicita acesso aos estudos e aguarda os provedores de saúde enviarem. Também é descrita a comunicação dos provedores de saúde (que armazenam as imagens) com o <i>framework</i> proposto, com o objetivo de prover privacidade aos pacientes. Ademais, no campo em vermelho a figura apresenta um cenário de ataque de re-identificação, em que um usuário malicioso tenta inferir a identidade dos pacientes a partir de dados obtidos dos médicos e pesquisadores. Por fim, a partir da proposta deste projeto para o médico a privacidade é protegida pelo método K-Anonimato e para o pesquisador com a Privacidade Diferencial.	85

Figura 8 – Resultados das simulações para a quantidade de CPU em percentagem que a blockchain consome por organização com diferentes valores de transações por segundo (TPS).	87
Figura 9 – Resultado das simulações para a quantidade de memória que a blockchain consome por organização com diferentes valores de transações por segundo (TPS).	88
Figura 10 – Resultado das simulações para o desempenho da rede para as operações de adição e consulta, considerando as métricas de latência e vazão.	89
Figura 11 – Influência da escolha do ϵ para as métricas entropia e utilidade. A métrica entropia é representada pela curva laranja e a utilidade pela curva verde, que são descritas seguindo a quantidade (eixo y) pelo valor do ϵ (eixo x).	94
Figura 12 – <i>Boxplot</i> para as quantidade de entropia de cada modelo de privacidade. Assim, o eixo x descreve os modelos utilizados, o K-Anonimato e Privacidade Diferencial e o eixo y apresenta a quantidade de entropia $H(X)$ normalizada.	95
Figura 13 – <i>Boxplot</i> para a utilidade dos dados extraído por cada modelo de privacidade, dado pela fórmula 1 menos a distância de cossenos. Assim, o eixo x descreve os modelos utilizados, o K-Anonimato e Privacidade Diferencial e o eixo y apresenta a dissimilaridade normalizadas.	97
Figura 14 – Caracterização das métricas de privacidade para a função densidade de probabilidades. O eixo x para a Figura 14(a) representa as distâncias com base na fórmula da dissimilaridade. Para a Figura 14(b) o eixo x representa a entropia, isto é, o nível de privacidade. O eixo y para ambas as figuras representam a densidade, ou seja, a probabilidade de concentração dos valores.	98
Figura 15 – Caracterização das funções densidade de probabilidade acumulada CDFs para a variável entropia, com representação para cada método de privacidade, K-anonimato e Privacidade diferencial.	99
Figura 16 – Caracterização das funções densidade de probabilidade acumulada (CDFs), para a variável utilidade, com representação para cada método de privacidade, K-anonimato e Privacidade diferencial.	100
Figura 17 – Diagrama de sequência que descreve o fluxo para o módulo de inserção do <i>token</i> no metadado da imagem DICOM.	123
Figura 18 – Diagrama de sequência que descreve o fluxo para auditar imagens vazadas com base no <i>token</i> contido nos metadados.	124

LISTA DE ALGORITMOS

Algoritmo 1 – Algoritmo Privacidade Diferencial modificado	80
Algoritmo 2 – Algoritmo K-Anonimato modificado	82

LISTA DE TABELAS

Tabela 1	– Número de trabalhos associados com as palavras-chave utilizadas na revisão bibliográfica, indicando uma sequência de estudos a partir das palavras-chave gerais para as específicas.	52
Tabela 2	– Sumário das características levantadas dos trabalhos relacionados, que foram usadas para comparar como a nossa solução.	66
Tabela 3	– Requisitos dos sistemas para comparar as propostas que visam prover privacidade aos pacientes que utilizam a estrutura da blockchain em sistemas de saúde. Adaptado de Shi <i>et al.</i> (2020).	68
Tabela 4	– Políticas de controle de acesso para os métodos de privacidade e componentes do <i>framework</i> . Também apresenta os usuários que compõem a arquitetura e as ações que os mesmos podem realizar.	79
Tabela 5	– Desempenho da rede quando se aplicam os módulos do provedor de saúde e de privacidade em conjunto com a blockchain. Foram consideradas as métricas, latência, vazão, <i>bytes</i> enviados e tempo, no qual as duas primeiras foram para avaliar o custo geral da rede comunicando com a blockchain e, a última, o tempo, foi para a verificar o custo de tempo de anonimização para cada método de privacidade.	90
Tabela 6	– Proximidade das funções densidade de probabilidade acumulativa para os valores reais de entropia a partir do teste <i>Komogorov-Smirnov</i> (K-S).	98
Tabela 7	– Proximidade das funções densidade de probabilidade acumulativa para os valores reais da utilidade, a partir do teste <i>Komogorov-Smirnov</i> (K-S).	99

LISTA DE ABREVIATURAS E SIGLAS

ANPD	Autoridade Nacional de Proteção de Dados
API	<i>Application programming interface</i>
CA	<i>Certification Authorities</i>
CBA	Criptografia Baseada em Atributos
CDF	<i>Culmulative Density Function</i>
CH	Criptografia homomórfica
DF	Privacidade Diferencial
DLT	<i>Distribuited Ledger Technology</i>
GCP	<i>Google Cloud Platform</i>
HLF	Hyperledger Fabric
IBM	<i>International Business Machines Corporation</i>
IPFS	<i>InterPlanetary File System</i>
JSON	<i>JavaScript Object Notation</i>
K-S	<i>Komogorov-Smirnov</i>
KA	K-Anonimato
MSP	<i>Membership Provider</i>
MX	<i>Mixing</i>
OSN	<i>Ordering Service Nodes</i>
OWASP	<i>Open Web Application Security Project</i>
PbD	<i>Privacy by Design</i>
PBFT	<i>Pratical Bizantine Fault Tolerance</i>
PCZNI	Prova de Conhecimento Zero Não interativa
PDF	<i>Probability Density Function</i>
PoS	<i>Proof-of-Stake</i>
PoW	<i>Proof-of-Work</i>
REM	Registros Eletrônicos Médicos
REST	<i>Representational state transfer</i>
SDK	<i>Software development kit</i>
TCIA	<i>The Cancer Image Archieve</i>
TPS	Transações por Segundo

SUMÁRIO

1	INTRODUÇÃO	25
1.1	Motivação	26
1.2	Objetivos	28
1.3	Organização do trabalho	28
2	FUNDAMENTAÇÃO TEÓRICA	31
2.1	Tecnologia blockchain	31
2.1.1	<i>Definições preliminares</i>	32
2.1.2	<i>Características do blockchain</i>	33
2.1.3	<i>Modelos de redes blockchain</i>	35
2.1.4	<i>Mecanismos de consenso</i>	36
2.1.5	<i>Plataforma Hyperledger Fabric</i>	38
2.2	Blockchain aplicado a saúde	40
2.3	Definições sobre privacidade	42
2.3.1	<i>Lei geral de proteção de dados pessoais</i>	42
2.3.2	<i>Modelos para garantir privacidade</i>	44
3	METODOLOGIA E TÉCNICAS DE PESQUISA	51
3.1	Levantamento bibliográfico	51
3.2	Formulação das questões de pesquisa	52
3.3	Seleção da estrutura de dados a partir da tecnologia de livro-razão distribuído	53
3.4	Seleção do modelo de rede blockchain	54
3.5	Seleção do protocolo de consenso e plataforma	56
3.6	Método de avaliação experimental	57
3.7	Considerações finais	59
4	TRABALHOS RELACIONADOS	61
4.1	Compartilhamento de informações de saúde baseado em blockchain	61
4.2	Preservação da privacidade em ambientes blockchain para saúde	63
4.3	Sumário dos trabalhos	65
4.3.1	<i>Visão geral</i>	65
4.3.2	<i>Discussão</i>	68

5	FRAMEWORK PROPOSTO PARA A PRESERVAÇÃO DA PRIVACIDADE	71
5.1	<i>Framework proposto</i>	71
5.1.1	<i>Descrição dos componentes do framework</i>	72
5.1.2	<i>Descrição do fluxo de dados do framework</i>	75
5.2	<i>Framework proposto frente a arquitetura de referência</i>	76
5.3	<i>Controle de acesso</i>	78
5.4	<i>Modelo de privacidade</i>	79
6	EXPERIMENTOS E RESULTADOS	83
6.1	<i>Configuração dos experimentos</i>	83
6.2	<i>Descrição do estudo piloto e cenário de ataque</i>	84
6.3	<i>Avaliação de desempenho</i>	86
6.3.1	<i>Avaliação da blockchain</i>	86
6.3.2	<i>Avaliação do módulo do provedor de saúde e de privacidade</i>	90
6.4	<i>Avaliação dos requisitos de privacidade</i>	91
6.4.1	<i>Avaliação qualitativa da privacidade</i>	91
6.4.2	<i>Avaliação quantitativa da privacidade</i>	93
6.5	<i>Análise teórica do token de rastreabilidade</i>	100
6.6	<i>Discussão</i>	102
7	CONCLUSÃO	105
7.1	<i>Principais contribuições</i>	106
7.2	<i>Produção científica</i>	107
7.3	<i>Trabalhos futuros</i>	107
	REFERÊNCIAS	109
	GLOSSÁRIO	121
APÊNDICE A	DIAGRAMA DE SEQUÊNCIA PARA O FLUXO DO TOKEN DE RASTREAMENTO	123

INTRODUÇÃO

Com os avanços tecnológicos e a integração da computação em diversas áreas do conhecimento, como a saúde, surgem diferentes soluções para potencializar os serviços convencionais (RAGHUPATHI; RAGHUPATHI, 2014; MIAH; GAMMACK; HASAN, 2019). Organizações privadas e governamentais estão projetando aplicações de saúde visando a adição de novas tecnologias para apoiar a tomada de decisão dos profissionais (MIAH; GAMMACK; HASAN, 2019). Algumas das tecnologias que podem trazer melhorias para a área são: Internet das Coisas (IoT), Inteligência Artificial e blockchain (RAGHUPATHI; RAGHUPATHI, 2014). Em paralelo com a integração de novas tecnologias também cresce a quantidade de dados gerados pelos sistemas de saúde (RAGHUPATHI; RAGHUPATHI, 2014). Grande parte desses dados são provenientes de registros eletrônicos médicos em hospitais, que podem ser exames, prontuários, imagens de radiografia e outros (BOONSTRA; VERSLUIS; VOS, 2014). Esses dados podem ser compartilhados entre diferentes centros de pesquisa em saúde, para aumentar a quantidade de dados e diversificar a gama de amostras existentes (BANZI *et al.*, 2019). A troca de informação entre pesquisadores é importante na descoberta de novas soluções (BANZI *et al.*, 2019). Um fato atual que mostra a importância no compartilhamento de dados de saúde é a pandemia do novo Coronavírus, em que diversos centros de pesquisas estão trocando informações com um único objetivo, desenvolver uma vacina eficaz contra o vírus (HOLUB *et al.*, 2020).

Porém, o compartilhamento de dados de saúde por parte das organizações deve ser uma prática cautelosa. Haja a vista que as aplicações na área da saúde, segundo um levantamento da IBM-X, são uma das que mais sofrem com ataques – resultando no vazamento de dados – que vão desde ataques de *Ransomware* e *Phishing*, até os re-identificação (CHEN, 2019). Por exemplo, em outubro de 2019 ocorreu o vazamento de 68 mil registros médicos no hospital metodista em Indiana nos Estados Unidos. Também no início de 2020 ocorreu o vazamento de cerca de 50 mil registros de pacientes no Minnesota hospital¹. Em 2019, segundo Seh *et al.*

---

¹ <https://portswigger.net/daily-swig/data-breach-at-minnesota-hospital-threatens-nearly-50-000-healthcare-records>

(2020) foram 41 milhões de usuários expostos de provedores de saúde, com o crescimento de 80% (de 2015 a 2019) no número de casos; enquanto outros serviços, como governamentais, sofreram crescimento de 20%. Diante disso, os sistemas de saúde estão suscetíveis a ataques de re-identificação, além de vazamentos em geral – ataques de re-identificação têm o propósito de re-identificar o paciente com base em um conhecimento prévio dos atributos de outra base de dados (GARFINKEL, 2015; KHAN; HOQUE, 2019).

Entre as tecnologias aplicadas para potencializar a área da saúde e melhorar os aspectos da confiança entre os nós, uma delas é a blockchain. Essa é uma tecnologia de livro-razão distribuído, que tem como principais características o armazenamento descentralizado e distribuído, a imutabilidade e a dispensa de um terceiro confiável (AL-JAROUDI; MOHAMED, 2019). Desde o seu surgimento, a blockchain tem passado por diversas transformações, até chegar em meados de 2018, que apresenta a sua aplicação em diversas áreas do conhecimento, como artes, finanças, governo e saúde (SWAN, 2015; ANDERSON, 2018; AZARIA *et al.*, 2016; LINN; KOO, 2016). Aplicado aos sistemas de saúde, pode prover maior confiabilidade no armazenamento de registros médicos e reduzir riscos de adulteração, proporcionando o não repúdio (SHAHNAZ; QAMAR; KHALID, 2019).

Mesmo com a integração da tecnologia blockchain para mitigar problemas relativos à confiança nos ambientes de saúde, ainda surgem problemas com a privacidade dos usuários que utilizam desses sistemas (AXON; GOLDSMITH; CREESE, 2018). Por exemplo, alguns identificadores como número de telefone, idade e código postal, podem ser combinados e utilizados por criminosos na aplicação de golpes, para, assim, tirar vantagens dos usuários. Essas preocupações se tonam mais críticas para as organizações, uma vez que os direitos dos usuários são protegidos por leis, tais como a Lei Geral de Proteção de dados (LGPD) (Lei n. 13.709/2018²), *General Data Protection Regulation* (GDPR) e *Health Insurance Portability and Accountability Act* (HIPAA) (FLAUMENHAFT; BEN-ASSULI, 2018). Nesse cenário, há necessidade da aplicação sistemática de políticas para o tratamento de dados por parte do agente de tratamento, nesse caso, principalmente as organizações de saúde. Sistemas de saúde podem se beneficiar com a aplicação da blockchain, no entanto, é preciso ajustá-lo, de forma que a rede tenha um conjunto de validadores confiáveis e ofereça um nível de privacidade aderente com a LGPD (SWAN, 2015; Jin *et al.*, 2019; FENG *et al.*, 2019).

1.1 Motivação

Motivado por riscos à segurança dos sistemas de saúde, a integração da tecnologia blockchain pode contribuir para mitigar esses problemas, principalmente para garantir maior confiança no compartilhamento dos dados, provendo o não repúdio e *logs* imutáveis (Jin *et al.*, 2019). No entanto, mesmo essa tecnologia auxiliando no contexto da confiança, ela ainda

² <http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2018/Lei/L13709.htm>

apresenta problemas referentes à privacidade, que são preocupações que devem ser consideradas pelos desenvolvedores dos sistemas blockchain, inclusive, quando aplicados aos sistemas de saúde (FENG *et al.*, 2019; AXON; GOLDSMITH; CREESE, 2018). Segundo Feng *et al.* (2019), os desafios da blockchain com privacidade são classificados em dois tipos: (i) privacidade de transação; e (ii) privacidade de identidade. Portanto, é necessário o estabelecimento de técnicas para reduzir os riscos com as transações e identidade dos usuários que partilham seus dados (FENG *et al.*, 2019).

Além disso, um problema que os sistemas de saúde e seus usuários devem se preocupar são os ataques de re-identificação, uma vez que estes têm o objetivo de descobrir a identidade do usuário com base em seus dados públicos, sejam de redes sociais ou de outro banco de dados de saúde (GARFINKEL, 2015). Na classe dos ataques de re-identificação, um deles é o ataque de ligação de registros, que associa registros públicos em outros bancos, que combinados indicam padrões de comportamento do usuário (KHAN; HOQUE, 2019). Para mitigar esse problema são utilizados métodos de anonimização, com o objetivo de eliminar a ligação com a identidade do usuário, entre eles estão o K-Anonimato (KA) e a Privacidade Diferencial (PD) (GARFINKEL, 2015). Utilizando essas técnicas para de-identificação, é possível reduzir os riscos de violação da privacidade.

Alguns trabalhos na literatura, como Guo *et al.* (2018), Al-Omar *et al.* (2019) e Gan *et al.* (2020) propõem abordagens para mitigar os problemas de privacidade em sistemas de saúde baseados em blockchain. O trabalho de Guo *et al.* (2018) apresentam uma proposta teórica para prover privacidade de transações a partir da criptografia baseada em atributos, mas, não considera riscos com ataques que expõem a identidade dos pacientes. Al-Omar *et al.* (2019), por sua vez, discutem uma proposta com uma blockchain pública para resolver o problema da privacidade das transações, auxiliada pela criptografia baseada em atributos, no entanto, ainda há problemas com a proteção da identidade dos usuários. Gan *et al.* (2020) oferecem uma estrutura de blockchain do Ethereum (BUTERIN, 2014), que garante a privacidade com base no controle de acesso para os ativos, porém, não se preocupam com ataques de ligação e com a identidade do usuário envolvido nas transações. Vale ressaltar que, segundo Garfinkel (2015) e Feng *et al.* (2019) os sistemas baseados em blockchain apresentam desafios relacionados a privacidade que são sobre: a identidade dos usuários, as transações e os ataques de re-identificação.

Este trabalho, motivado pelos problemas de privacidade de identidade e das transação, assim como pelos riscos aos ataques de re-identificação, aos quais os sistemas blockchain para saúde estão suscetíveis. Esta pesquisa tem o propósito de explorar este tema, tendo como base a seguinte questão de pesquisa:

“Como mitigar os riscos à privacidade dos pacientes no compartilhamento de dados de saúde em redes baseadas em blockchain?”

1.2 Objetivos

Norteados pela questão de pesquisa este trabalho teve como objetivo desenvolver um *framework* baseado em blockchain que preserve a privacidade no compartilhamento de dados de saúde e reduza de riscos a ataques de re-identificação. Assim, foram utilizados modelos de anonimização — os modelos K-Anonimato e Privacidade Diferencial (ver na Seção 2.3.2) — para prover a privacidade dos pacientes quando estes compartilham seus dados com pesquisadores e profissionais de saúde. Para complementar esse objetivo principal, há os seguintes objetivos específicos:

- Escolher a estrutura de livro-razão distribuído mais adequada para esta proposta;
- Selecionar o tipo de rede e processo de consenso que melhor se adapta ao cenário dos sistemas de saúde;
- Selecionar os modelos de privacidade mais adequados para mitigar os riscos a privacidade em sistemas de saúde baseados em blockchain;
- Implementar o modelo de privacidade e integrar à rede blockchain;
- Mensurar o desempenho de rede e hardware do *framework*, além de analisar as despesas que pode causar aos sistemas de saúde;
- Rastrear os ativos compartilhados entre os pesquisadores com auxílio da blockchain;
- Mensurar e analisar a proposta, com base em métricas estatísticas, além da efetividade dos modelos de privacidade integrados ao *framework*.

Ademais, para este projeto foi elaborado um estudo piloto para o compartilhamento de imagens médicas no formato *Digital Imaging and Communications in Medicine* (DICOM), que emprega a blockchain para armazenamento de *logs* e metadados da imagem. Assim, os arquivos de imagens DICOM são armazenados localmente nas bases de dados dos provedores de saúde, sendo ligados com a rede *Off-chain* do *InterPlanetary File System* (IPFS) para os usuários acessarem de forma descentralizada.

1.3 Organização do trabalho

Este documento foi organizado da seguinte forma. O Capítulo 2 exibe o referencial teórico e os conceitos que foram empregados neste trabalho. O Capítulo 3 descreve como a pesquisa foi conduzida e quais foram os métodos utilizados para levantar os trabalhos e responder às questões propostas. O Capítulo 4 retrata os trabalhos da literatura que estão associados aos temas blockchain, saúde e privacidade, para, assim, levantar os problemas da área. O Capítulo 5 expõe como o *framework* proposto foi organizado e quais são seus componentes, bem como seu

processo de implementação. O Capítulo 6 descreve como os experimentos foram planejados e realizados, no cenário de estudo do compartilhamento de imagens médicas, além de apresentar os resultados dos testes de desempenho e privacidade para os métodos, tal como para a blockchain. Por fim, o Capítulo 7 apresenta as conclusões sobre esta pesquisa, além de explicitar as principais contribuições, trabalhos futuros e produções científicas resultantes.

FUNDAMENTAÇÃO TEÓRICA

Esta seção tem o propósito de apresentar os conceitos, técnicas e ferramentas que foram abordados nesta pesquisa e estão presentes nas próximas seções do trabalho.

2.1 Tecnologia blockchain

A tecnologia de livro-razão distribuído (do inglês, *Distributed Ledger Technology* (DLT)) vem ganhando visibilidade nos últimos anos, devido à ascensão das criptomoedas. Essa tecnologia pode potencializar outros sistemas, além do financeiro, com as moedas digitais. Assim, a blockchain pode acrescentar a tais tecnologias, os contratos inteligentes e uma estrutura imutável baseada em blocos interligados para prover maior segurança aos sistemas computacionais (SUN; HUANG, 2018; SWAN, 2015). Além disso, o blockchain provê uma arquitetura descentralizada, que não necessita de uma entidade central confiável (como o banco central) para gerenciar as transações, isto é, evita que uma terceira parte acesse as informações dos usuários (BODKHE *et al.*, 2020).

A primeira arquitetura blockchain surge com a criptomoeda Bitcoin, que possibilita aos usuários realizarem transações financeiras de forma pseudo-anônima na internet, sem necessitar de cadastro em uma agência intermediadora. Com uma estrutura de blocos ligados por um ponteiro *hash* e com um livro-razão distribuído e replicado entre os pares da rede, esse modelo pode prover maior confiabilidade na troca de informações (SUN; HUANG, 2018; SWAN, 2015). Além da estrutura de blocos, outras surgem para resolver os problemas relacionados ao alto custo computacional no processo de mineração, como o Tangle proposto por Popov (2018), que segue uma estrutura baseada em grafos para armazenar as transações (KUSMIERZ; STAUPÉ; GAL, 2018). Desse modo, esta seção tem como objetivo apresentar os conceitos básicos sobre blockchain, que é a primeira a surgir e será empregada neste trabalho. A seguir serão apresentadas as principais definições que compõem esta tecnologia.

2.1.1 Definições preliminares

Esta seção apresenta definições preliminares importantes para o entendimento dos de blockchain. São elas:

- **Criptografia de chave pública:** é um método de criptografia assimétrica que utiliza duas chaves (pública e privada) para encriptar e decriptar os dados trocados entre duas pontas, que podem ser dois *smartphones* trocando mensagens no *WhatsApp* (STALLINGS, 2014);
- **Hash:** é um valor gerado por uma função que recebe uma mensagem arbitrária m e produz um valor $h = H(m)$ de tamanho fixo (STALLINGS, 2014). Além disso, pode ser uma função criptográfica de mão única, ou seja, não tem função inversa. Nesse sentido, o valor original, se alterado, muda totalmente o *hash* de saída. Outra propriedade, define difícil a tarefa de encontrar um valor arbitrário que seja idêntico a um *hash* já gerado (STALLINGS, 2014);
- **Assinatura digital:** com esse mecanismo é possível obter um código único para utilizá-lo como assinatura. Na composição da assinatura utiliza-se da criptografia de chave pública para criptografar um valor *hash* e assinar documentos digitalmente entre duas pontas. Desse modo, um cliente utiliza da chave privada para assinar o documento e outro utiliza da chave pública para fazer a leitura (STALLINGS, 2014);
- **Redes Peer-to-peer (P2P):** é uma arquitetura de sistemas de computação distribuída, de maneira que as máquinas cooperem entre si, exercendo os papéis tanto de cliente quanto de servidor (COULOURIS *et al.*, 2013);
- **Mineração e validação:** é um processo previsto na estrutura blockchain do Bitcoin, que tem a função de verificar a veracidade das transações enviadas para rede. Esse processo possui alto custo computacional, pois executa sucessivas tentativas para descobrir um valor *hash* que siga o padrão estabelecido pela rede, de modo a validar o bloco enviado (XU *et al.*, 2017);
- **Nonce:** é um valor que fica no cabeçalho do bloco, sendo adicionado na função para gerar o *hash* do bloco e, conectado aos demais atributos. Esse valor é incrementado até que se encontre um *hash* que represente o *hash* do bloco concatenado com bits zeros no início (NAKAMOTO, 2008);
- **Transação:** são registros da sequência de operações realizadas entre os clientes de um sistema, que pode ser a troca de um ativo importante para as partes (YAGA *et al.*, 2018);
- **Bloco:** é o principal elemento da estrutura de dados da blockchain, para armazenar o cabeçalho e as transações. O cabeçalho é composto pelos seguintes atributos (ZHENG *et al.*, 2018): o *hash* de referência da árvore de Markle; o *hash* do bloco anterior; a versão do

bloco; o número de bits; o *Timestamp*; e o Nonce. Todo o bloco novo é ligado ao anterior por um ponteiro *hash* para tornar os dados imutáveis (HAFFKE, 2017);

- **Token:** é um objeto virtual que identifica um ativo, indicando a unidade de valor a ser transferida na estrutura do blockchain. Por exemplo, a moeda no Bitcoin e o Ethereum são classificados como *token* (HAFFKE, 2017).

2.1.2 Características do blockchain

Primeiramente, voltando à característica geral do blockchain, temos sua invenção, proposta por Nakamoto (2008), que estabelece um sistema monetário eletrônico baseado em P2P, ou seja, a criptomoeda Bitcoin (GUO *et al.*, 2018). A partir dos avanços desse achado pode-se definir formalmente o blockchain como uma tecnologia de livro-razão distribuído, alicerçada em uma rede P2P e tendo como principal componente um livro que registra as transações de forma pública ou privada em ordem cronológica (MUKHOPADHYAY *et al.*, 2016). Para formar a estrutura, os blocos são ligados entre si, a partir do ponteiro *hash* do bloco anterior (NAOR; YUNG, 1989), possibilitando que os dados do bloco não sejam adulterados. Outra definição dessa tecnologia, segundo Rifi *et al.* (2017), é uma arquitetura de computação distribuída baseada em P2P, com um livro-razão replicado e distribuído. Assim, a Figura 1 representa a estrutura de dados do blockchain, em que cada bloco é ligado ao anterior por ponteiros criptográficos (*hash*) até o bloco gênese (primeiro bloco da estrutura) e as transações são agrupadas dentro dos blocos (DINH *et al.*, 2018).

Os avanços da tecnologia blockchain, tem sido aplicados aos ativos monetários, tal como em diferentes áreas, como saúde, agricultura e outras. Segundo Swan (2015) e Hoy (2017), a blockchain pode ser categorizada em três fases. A blockchain 1.0 é caracterizada pelo surgimento das criptomoedas com Nakamoto (2008), bem como pela sua implantação no mercado, com a transferência de ativos simples. A blockchain 2.0, por sua vez, é caracterizada pelo surgimento dos contratos inteligentes (do inglês, *Smart Contracts*), com base no artigo de Buterin (2014), e pelo avanço na forma de se trabalhar, saltando de transações simples para transações mais complexas. Já a blockchain 3.0 é a evolução que estamos observando atualmente, com a aplicação dessa tecnologia não somente em finanças, mas em diferentes cenários, tais como medicina, ciência, agricultura e cadeias de suprimentos.

Portanto, algumas características da blockchain que podem contribuir para potencializar os sistemas convencionais são as seguintes (GREVE *et al.*, 2018; LIN; LIAO, 2017; XU *et al.*, 2017):

- **Descentralização:** é uma característica importante, sendo um diferencial comparado a outras tecnologias na indústria, pois as aplicações convencionais necessitam que uma terceira parte faça a interação com as transações dos usuários. Os sistemas baseados em blockchain não necessitam de um nó centralizado confiável para tratar as operações,

Figura 1 – Estrutura de dados do blockchain. Adaptado de (DINH *et al.*, 2018).

tudo é feito de maneira distribuída e descentralizada. Assim, reduz problemas como a indisponibilidade do nó central, uma vez que, cópias dos dados são disseminadas pela rede;

- **Transparência:** as transações são registradas em um livro-razão público e são auditáveis pelos nós da rede. No entanto, o nível de transparência depende da classificação de rede utilizada, como permissionada ou não-permissionada, e do desempenho da aplicação;
- **Imutabilidade:** as transações registradas no livro-razão não podem ser adulteradas, a menos que uma entidade consiga tomar controle de 50 % dos nós. Essa propriedade também está diretamente associada ao ponteiro *hash* que interliga os blocos tornando-os livres de fraudes, uma vez que, se houver alguma alteração nos dados do bloco o *hash* referente ao bloco muda completamente, podendo ser identificado pela rede;
- **Pseudo-privacidade:** é possível oferecer privacidade aos pares, desde que uma transação seja executada de maneira pseudo-anônima, resolvendo o problema de confiança entre os pares, visto que, o rastreamento dos ativos é complexo, como vemos no Bitcoin. Cada usuário faz o controle de suas próprias chaves e tem a autoridade para disponibilizá-las a quem desejar;

- **Incentivo:** Algumas aplicações disponibilizam um incentivo aos usuários da rede, apoiado por um desafio que deve ser solucionado, por exemplo, no caso do Bitcoin o incentivo é uma quantia de criptomoedas para o vencedor. Contudo, os sistemas podem definir diferentes modelos de negócios para essas vantagens, podendo até mesmo se basear em propriedades da teoria dos jogos. Vale ressaltar que, o incentivo não está necessariamente presente em todas as estruturas blockchain. Assim, o mecanismo de incentivo não necessita ser financeiro, pode ser integrado de outras formas;
- **Consenso:** é um mecanismo para estabelecer acordo entre os nós da rede sobre a condição de uma operação, para que seja validada ou rejeitada. Na blockchain alguns algoritmos que estabelecem o consenso são: Proof-of-Work (PoW), baseado em falhas bizantino, Proof-of-Stake (PoS) e outros (YAGA *et al.*, 2018);
- **Confiança:** a blockchain fornece aos usuários, maior confiança na troca de ativos, visto que é imutável, auditável, transparente e utiliza de um protocolo para validar os blocos. Desse modo, com base no protocolo de consenso e na interação entre os nós, é estabelecido um acordo para examinar a veracidade dos blocos inseridos, para, assim, ter uma troca de ativos honesta;
- **Contratos inteligentes:** os contratos inteligentes partem da ideia de propriedades inteligentes para a blockchain. O objetivo é possibilitar aos sistemas algo a mais do que simples transações de trocas, estabelecendo acordos dinâmicos entre as partes. Desse modo, proporcionando maior confiança na negociação de troca de ativos digitais (SWAN, 2015).

Por fim, fundamentada nessas características a tecnologia blockchain foi construída e incrementada, apresentando potencial de aplicação em áreas multidisciplinares, como é o caso das aplicações em saúde, que é tema desta pesquisa e serão discutidas com mais detalhes nas próximas seções.

2.1.3 Modelos de redes blockchain

A blockchain pode ser caracterizada em dois grupos, conforme a forma como os dados são protegidos. São eles (WUST; GERVAIS, 2018; ALHADHRAMI *et al.*, 2017): permissionado (do inglês, *permissioned*) e não-permissionado (do inglês, *permissionless*). Além disso, a permissionado pode ser público ou privado. Assim, as definições para as redes podem ser categorizadas da seguinte forma:

- **Permissionado público:** o processo de ingresso na rede é controlado, necessita-se de autenticação, assim, somente os nós com acesso permitido poderão ingressar na rede. Para estabelecer o consenso, os membros autenticados podem fazer parte do conjunto de nós

validadores. A leitura pode ser feita pelos membros autenticados, ou ser restrita somente para nós especificados (ALHADHRAMI *et al.*, 2017);

- **Permissionado privado:** os nós, também necessitam se autenticar para ingressar na rede, no qual as permissões para persistir os dados e manter a rede, são centralizadas em uma ou várias organizações confiáveis, que tem a função de adicionar o novo bloco à cadeia, após a validação. O processo de leitura pode ser feito de forma aberta para os nós autenticados ou ser restrito para um conjunto confiável (ALHADHRAMI *et al.*, 2017);
- **Não-permissionado:** são redes blockchain com acesso aberto sem a necessidade de autenticação, ou seja, qualquer usuário pode criar uma carteira e utilizá-la. É composta por um livro-razão transparente, em que qualquer par da rede pode se tornar um minerador e tentar validar os blocos, como no caso do Bitcoin e do Ethereum (WUST; GERVAIS, 2018).

Em suma, a divisão da blockchain em categorias é importante para possibilitar a customização das aplicações, com diferentes regras de acesso, a fim de sistematizar a organização dos modelos de rede com base nas necessidades da aplicação. Quando uma rede blockchain é projetada, o modelo de rede selecionado, se for inadequado, pode comprometer a privacidade do usuário ou os dados podem ficar muito restritos.

2.1.4 Mecanismos de consenso

Os protocolos de consenso são aplicados aos sistemas computacionais distribuídos e fundamentados pelo problema da confiança entre os pares (LAMPORT; SHOSTAK; PEASE, 1982). Assim, Lamport, Shostak e Pease (1982) apresentaram o problema dos **Generais Bizantinos**, que ilustra uma complicação sobre a confiabilidade da informação transferida entre os pares participantes da rede (BACH; MISAKJEVIC; ZAGAR, 2018). Essa analogia é descrita da seguinte forma: n generais devem atacar uma fortaleza, contudo, é necessário que todos ataquem simultaneamente, para que a fortaleza seja derrubada. Desse modo, para atacar de forma sincronizada, os generais trocam mensagens de alerta entre si, com o propósito de informar o momento exato do ataque e ter maior assertividade. No entanto, entre os generais podem existir traidores que não queiram atacar, ou até mesmo que tentem fraudar as mensagens enviadas para os outros generais (LAMPORT; SHOSTAK; PEASE, 1982). Sendo assim, os protocolos de consenso têm o objetivo definir regras de convergência para uma solução comum (BACH; MISAKJEVIC; ZAGAR, 2018).

Consequentemente, por se tratar de uma estrutura descentralizada, é necessária confiança entre os pares em ambientes blockchain, logo, são aplicados os protocolos de consenso para estabelecer confiança na validação dos blocos que serão adicionados na cadeia (CACHIN; VUKOLIC, 2017). Esses protocolos quando aplicados ao blockchain utilizam dos próprios nós

da rede como validadores e seguem uma regra pré-estabelecida para verificar se as mensagens trocadas são verdadeiras (CACHIN; VUKOLIC, 2017). Alguns dos protocolos de consenso utilizados nas plataformas mais famosas são: (i) *Proof-of-Work* (PoW), utilizado pelo Bitcoin (NAKAMOTO, 2008); (ii) *Practical Byzantine Fault Tolerance* (PBFT), utilizado pelo Hyperledger Fabric (ANDROULAKI *et al.*, 2018); (iii) *Proof-of-Stake* (PoS), empregado no Cardano (GAŽI; KIAYIAS; ZINDROS, 2019); (iv) Tangle, utilizado pelo IOTA (POPOV, 2018). Os próximos parágrafos descrevem os fundamentos desses protocolos de consenso.

O protocolo **Prova de trabalho** (do inglês, *Proof-of-Work*) é utilizado pelo Bitcoin e tem como principal características a competição entre os nós para resolver um enigma matemático. Esse processo resulta em uma recompensa monetária, dada ao primeiro que resolve o problema (MINGXIAO *et al.*, 2017). Para o funcionamento desse protocolo é lançado um enigma criptográfico para os pares da rede resolverem, isso é chamado de mineração no blockchain do Bitcoin. Esse processo consiste essencialmente em diversas tentativas exigindo grande poder computacional para descobrir o enigma, que deve ser realizado sem ultrapassar o número máximo de iterações a partir da quantidade de zeros definidos no valor *hash* do bloco (MINGXIAO *et al.*, 2017). O nó que consegue resolver o enigma primeiro ganha o direito de inserir o bloco na rede e, assim, ganha uma recompensa (GREVE *et al.*, 2018; MINGXIAO *et al.*, 2017).

Outro protocolo viável, que será empregado nesta pesquisa é **baseado em Tolerância a Falhas Bizantinas Práticas** (do inglês, *Practical Byzantine Fault Tolerance*), que segue o problema dos generais Bizantinos, de modo a mitigar o recebimento ou envio de mensagens não confiáveis (CACHIN; VUKOLIC, 2017). Também vale ressaltar que este protocolo tem um limiar aceitável de nós que podem falhar (f), a partir do número total de nós (n), desde que siga a seguinte regra $f < n/3$ (CACHIN; VUKOLIC, 2017). O PBFT tem o objetivo de reduzir a complexidade do algoritmo, tornando-o mais leve, com base nos seguintes passos (MINGXIAO *et al.*, 2017):

1. É eleito um nó líder, a partir do conjunto de pares da rede;
2. Uma solicitação é enviada ao líder para registrar a data e hora de entrada do cliente;
3. Uma mensagem é pré-preparada pelo cliente, na qual o nó líder registra a solicitação e passa para um próximo nó verificar se ela é válida para ser enviada aos pares da rede ou não;
4. Se o nó aceitar a mensagem, é feita a transmissão para os outros pares verificarem a veracidade que, se aprovada, entra em estado de confirmação;
5. O líder aguarda a confirmação dos outros pares e se o número de mensagens confirmadas for $2 \times f + 1$, o líder julga as mensagens como autênticas;

6. Por fim, uma resposta é enviada para o cliente a partir do líder, indicando que a mensagem foi aprovada e está em estado de operação. Além disso, se a mensagem falhar é feito o reenvio dela.

O protocolo de consenso **Prova de participação** (do inglês, *Proof-of-stake*) utiliza do conceito de idade da moeda para deliberar os direitos e regras de acesso a determinados pares da rede (MINGXIAO *et al.*, 2017). Os usuários que obtiverem mais moedas e ficarem mais tempo com elas (moedas mais antigas) terão um custo menor para minerar (MINGXIAO *et al.*, 2017). Portanto, com esse processo pode-se reduzir ataques (como a tomada de 51% dos nós), pois os usuários maliciosos necessitam adquirir uma quantidade significativa de moedas e passar mais tempo com elas para conseguir tomar posse da rede (CAI *et al.*, 2018).

Por fim, um protocolo que segue outro padrão de validação, diferentemente dos citados, que são específicos para cadeia de blocos, é o **Tangle** que foi construído com base em uma estrutura de grafos (POPOV, 2018). A validação no Tangle é realizada por meio da ligação dos nós. Cada transação acrescentada é obrigada a validar duas outras transações, para, assim, tornar-se confiável (CACHIN; VUKOLIC, 2017). Quando uma transação é validada, um peso é atribuído para a aresta das outras transações que estão ligadas a ela, dessa forma, quanto mais a rede cresce, maior será esse peso e mais confiável serão as transações (CACHIN; VUKOLIC, 2017).

2.1.5 Plataforma Hyperledger Fabric

Concomitantemente aos avanços da tecnologia blockchain surgem plataformas para facilitar a construção dessas redes, auxiliando no desenvolvimento de redes customizadas e de contratos inteligentes (DANNEN, 2017). Uma das primeiras que surgiram foi o Ethereum, que é baseada na infraestrutura da moeda *Ether* e possibilita a construção de contratos inteligentes de forma simplificada, interligando-os a uma aplicação específica (BUTERIN, 2014). Além disso, a *International Business Machines Corporation* (IBM), em parceria com a *Linux Foundation*, criou o Hyperledger Fabric (HLF), viabilizando o desenvolvimento de redes permissionadas e contratos customizados. Desse modo, o HLF será discutido de forma detalhada a seguir, pois foi a plataforma empregada no desenvolvimento desta pesquisa e assegura o desenvolvimento personalizado de contratos inteligentes para as necessidades deste trabalho.

O HLF provê uma blockchain permissionada que trabalha com diversos protocolos de consenso, como o PBFT e o RAFT (ONGARO; OUSTERHOUT, 2014). Também apresenta uma estrutura com *broadcast* atômico, execução determinística e estado persistente em todos os pares da rede de forma sincronizada (ANDROULAKI *et al.*, 2018). Os principais componentes, formados com base em um conjunto de pares específicos, são: clientes, pares de execução (do inglês, *Peers* (P)), serviço de pedidos dos nós (do inglês, *Ordering Service Nodes* (OSN)) e provedor de associação (do inglês, *Membership Provider* (MSP)). A Figura 2 descreve a comunicação

entre cada componente da rede, que é gerenciado pelo MSP, cuja função é organizar a interação dos pares na rede e a distribuição dos certificados gerados pelas autoridades certificadoras (do inglês, *Certification Authorities* (CA)). Ademais, há pares de execução que são importantes, pois armazenam uma cópia do livro-razão e processam as propostas de transações enviadas pelos clientes. Já o OSN faz a ordenação dos pedidos enviados pelos pares de execução, para que sejam sincronizados e executados de forma coordenada. Por fim, para realizar a comunicação entre os pares de execução e os OSNs, um canal chamado *gossip* é estabelecido, que observa chegada de novas transações e encaminha para o OSN (ANDROULAKI *et al.*, 2018).

Figura 2 – Visão geral do *framework* Hyperledger Fabric. Adaptado de (ANDROULAKI *et al.*, 2018).

No que se refere à infraestrutura, o HLF emprega *containers* sustentados pelo Docker¹, composto por máquinas virtuais que consomem poucos recursos, tornando a implementação relativamente escalável e multiplataforma. Para iniciar o projeto com o Docker, o desenvolvedor obtém as imagens no site oficial, em que cada uma representa uma das abstrações citas anteriormente. As imagens virtuais necessárias para construir uma rede blockchain completa são: pares de execução, CA, OSN e pares para armazenamento (que é o banco de dados CouchDB²). Cada componente é associado a uma organização que tem vários pares de execução ligados diretamente aos pares de armazenamento para guardar as transações, além de utilizarem de um CA para ingressar na rede e obter certificados válidos (ANDROULAKI *et al.*, 2018).

No estabelecimento da comunicação no HLF podem ser configurados canais para a troca de ativos, que são sub-redes para o envio de transações privadas. Em um canal, a comunicação pode ser separada entre: os membros de uma organização; a organização completa; o livro-razão;

¹ <https://www.docker.com/>

² <https://couchdb.apache.org/>

e a aplicação que executa um contrato, ou os serviços de ordenação. Para o projeto da rede pode ser definido o número de canais e como será realizada a comunicação entre os membros de cada canal. Desse modo, antes iniciar a rede, alguns nós podem ser selecionados para compor os canais que forem mais adequados e, para ingressar precisam se autenticar. O conceito de canais no HLF é importante para separar as políticas de acesso e tornar a aquisição dos dados das organizações mais restritivo e seguro (FABRIC, 2020a).

Para tornar as transações dinâmicas, o HLF integra os contratos inteligentes, também chamados de *chaincode*, que estabelecem as regras de negócio para as transações da rede, podendo transferir qualquer ativo. Os contratos podem ser desenvolvidos em diferentes linguagens, como Golang, JavaScript, Java e TypeScript, fornecendo um *framework* completo para customizar o conteúdo e lógica das transações. Além disso, a ferramenta Visual Studio Code (VSCode) disponibiliza um *plugin* para facilitar o desenvolvimento e iniciar uma rede de teste. Ao final, para que os contratos se comuniquem com os pares da rede, é definido um ou vários canais que um conjunto de pares irá utilizar para enviar transações baseados nas regras do contrato (ANDROULAKI *et al.*, 2018).

Com relação à segurança, o HLF utiliza certificados digitais X.509, que são gerenciados pelos CAs e MSP, com o propósito de controlar os usuários que entram nas organizações e determinar quais recursos eles podem ter acesso (FABRIC, 2020d). Por exemplo, em um sistema de saúde, se um paciente tentar recuperar registros financeiros, que ficam na organização do provedor de saúde. Mesmo o paciente pertencendo a uma organização chamada hospital, seu certificado será inválido para obter esse recurso, pois, o MSP só concede acesso a alguns recursos do hospital, como seus próprios registros. Além disso, outro recurso sobre a segurança no HLF são as coleções privadas, que protegem os dados das organizações que estão armazenados no CouchDB, para que não sejam recuperados por usuários indesejados (FABRIC, 2020e). Para isso, o HLF cria um arquivo no formato JSON, que define as coleções e as categorias de usuários que irão recuperar cada um dos recursos (FABRIC, 2020e). Assim, é possível configurar algumas organizações que compartilham dados a partir de uma coleção privada, na qual somente os usuários da coleção conseguem ver os dados no livro-razão. Como exemplo, duas organizações, provedor de saúde e pesquisadores, criam uma coleção privada que contém atributos correspondentes a exames de sangue realizados pelos pacientes. Nesse caso, o protocolo de *gossip* permite que apenas os usuários das organizações provedor de saúde e pesquisadores recuperem os registros no blockchain (FABRIC, 2020e).

2.2 Blockchain aplicado a saúde

A área da saúde passou por diversas transformações, até alcançar a estrutura atual que segundo Bodkhe *et al.* (2020) é classificada como saúde 4.0. Essa fase da indústria da saúde se destaca pela aplicação de diferentes tecnologias, como Internet das Coisas, Aprendizado

de Máquina e blockchain. Com os avanços, também cresce o número de ataques cibernéticos sofridos pelos sistemas de saúde. Assim, a tecnologia blockchain tem potencial para auxiliar na redução dos efeitos causados pelos ataques (BODKHE *et al.*, 2020). Ademais, é possível apoiar os sistemas de saúde com base em uma estrutura descentralizada de armazenamento, com livro-razão imutável e confiável para proteger a integridade dos dados, com transações transparentes e um processo metódico de validação das transações para evitar violações (LI *et al.*, 2019; BODKHE *et al.*, 2020).

Aplicações de saúde utilizam do blockchain para o gerenciamento de registros médicos, cadeias de suprimentos para medicamentos e no monitoramento de pacientes (AGUIAR *et al.*, 2020). Esse uso pode proporcionar melhorias, por não necessitar de um terceiro confiável para gerenciar a troca de ativos, bem como por apresentar registros de *logs* imutável, garantindo o não repúdio. No entanto, a aplicação do blockchain nesse contexto apresenta algumas fraquezas, que novos estudos visam melhorar. São elas: alta complexidade para implementação, consumo elevado de energia e problemas com a privacidade dos pacientes (DUJAK; SAJTER, 2018; AGUIAR *et al.*, 2020).

Os usuários que utilizam dos sistemas de saúde tem preocupações constantes com a privacidade, pois esses sistemas tratam informações sensíveis, que exigem cuidados específicos. O sucesso dos sistemas de saúde está relacionado com a capacidade de proteção, uma vez que, a violação desses sistemas pode trazer grandes consequências para as organizações, como perdas financeiras e da reputação perante a sociedade (LIN; LIAO, 2017; TANWAR; PAREKH; EVANS, 2020). A blockchain pode auxiliar na melhoria da segurança na transferência dos dados dos pacientes (ZHANG; XUE; LIU, 2019). No entanto, mesmo a blockchain proporcionado melhorias em relação à confiabilidade dos sistemas de saúde, ainda apresenta desafios sobre privacidade (FENG *et al.*, 2019). Portanto, os sistemas de saúde baseados em blockchain apresentam alguns desafios em preservar a privacidade dos usuários, considerando dois requisitos. São eles: (i) a privacidade de transação; e (ii) a privacidade de identidade (FENG *et al.*, 2019). Dos quais são descritos da seguinte forma:

- **Privacidade de identidade:** está relacionado com a ligação da carteira digital com o usuário real, porque na blockchain é convencional a utilização de pseudônimos para se referir ao portador de uma determinada carteira digital. Ademais, os dados dos pacientes nas transações podem representar indicadores sobre a identidade deles, o que pode ocasionar que eles sejam re-identificados com base em técnicas específicas (FENG *et al.*, 2019);
- **Privacidade de transação:** está vinculada à problemática das transações só serem acessadas pelos nós com permissão de acesso. No entanto, na blockchain principalmente no não-permissionado, o conteúdo das transações que estão presentes no livro-razão podem ser validados por validadores desconhecidos. Desse modo, quando se trata de informações sensíveis é necessário um nível maior de privacidade, assim, surgem técnicas para mitigar

esse problema, como o blockchain permissionado (FENG *et al.*, 2019).

2.3 Definições sobre privacidade

A coleta de informações pessoais por instituições governamentais e não governamentais é imprescindível atualmente, pois pessoas geram um extensivo conjunto de dados de diferentes fontes (AIELLO *et al.*, 2019). Com a exposição dos dados para essas fontes (como redes sociais, sistemas de saúde, bancos e outros) cresce o número de casos de violação da privacidade, seja por invasões de hackers, ou até mesmo pelo gerenciamento indevido por parte da organização de tratamento (HUGHES, 2015). Com o crescimento dessa prática, surge a carência em proteger dados pessoais em sistemas eletrônicos. Em vista disso, surgem leis de privacidade que estabelecem os direitos e deveres dos agentes de tratamento perante aos dados de seus usuários. Algumas dessas leis são: *General Data Protection Regulation* (GDPR) da união Europeia em 2016, *Health Insurance Portability and Accountability Act* (HIPAA) dos Estados Unidos em 1996; e a Lei Geral de Proteção de Dados Pessoais (LGPD) do Brasil em 2020 (FLAUMENHAFT; BEN-ASSULI, 2018).

A proteção à privacidade dos indivíduos pode ser entendida como a garantia individual de uma pessoa natural ter seus dados preservados segundo a lei, para que não ocorra publicação, intervenção ou abusos com relação à utilização desses dados (LI *et al.*, 2019). Esse conceito é aplicado aos sistemas computacionais, para salvaguardar os dados dos usuários, que podem ser áudio, vídeo, texto, registros médicos e outros, de modo a realizar operações sob os dados, com consentimento dos proprietários (LI *et al.*, 2019). As operações podem ser realizadas sob um conjunto de dados pessoais que, segundo o artigo 4 da GDPR, são quaisquer dados que estão associados a uma pessoa física que pode ser identificada diretamente ou indiretamente (COMMISSION, 2018).

Outra definição importante que será utilizada neste trabalho, são sobre os dados pessoais sensíveis. Esses dados, segundo a LGPD, caracterizam a ligação a um indivíduo e representam origem racial ou étnica, crença religiosa, posição política, associação a sindicato ou grupo religioso, condições de saúde, expressão genética ou dado biométrico (BRASIL, 2018). No escopo deste trabalho, foram empregados dados de saúde que são sensíveis, assim a LGPD e GDPR aplicam medidas mais restritivas no tratamento. Segundo o artigo 11 da LGPD, o titular dos dados sensíveis deve fornecer consentimento de forma específica e clara, para dados de saúde os procedimentos de tratamento somente devem ser realizados por provedores de saúde e autoridades sanitárias (BRASIL, 2018).

2.3.1 Lei geral de proteção de dados pessoais

Como foi citado anteriormente, a LGPD surgiu da necessidade de regular o tratamento de dados pessoais, resultante de diversos casos de manipulação indevida. Inspirada na GDPR,

que entrou em vigência em 2018, proposta pela União Europeia, com o intuito de conceder aos indivíduos ferramentas legais para obrigar as organizações a fornecerem aos titulares detalhes sobre o processo de tratamento dos dados, além de fornecer medidas protetivas, para que o tratamento não cause danos a eles (Jin *et al.*, 2019; FLAUMENHAFT; BEN-ASSULI, 2018). A LGPD surgiu com o mesmo intuito da GDPR e com semelhanças em seus artigos, mas, adaptada ao contexto jurídico brasileiro, tal como para as necessidades específicas dos cidadãos do Brasil. Em vista disso, a LGPD será discutida com mais detalhes a seguir, logo, este estudo foi realizado em território brasileiro, sendo assim, esta lei que será válida.

A LGPD é fundamentada sob um órgão geral chamado Autoridade Nacional de Proteção de Dados (ANPD), que tem a função de regular todos os processos, verificar se os agentes de tratamento estão cumprindo as definições e aplicar punições quando ocorrer violações (BRASIL, 2018). O processo de tratamento de dados pessoais verificados pela ANPD inclui qualquer procedimento aplicado sob os dados, os quais podem ser: armazenamento, transferência, coleta, produção, classificação, processamento e eliminação (BRASIL, 2018).

Para realizar o tratamento dos dados segundo o artigo 4 da lei, são definidos alguns agentes para realizarem os procedimentos sob os dados, tais agentes podem ser uma pessoa física ou jurídica. Os agentes são: o controlador de dados que tem o direito de tomar as decisões a respeito do tratamento dos dados, e o operador de dados que executa o tratamento dos dados a partir das tarefas estabelecidas pelo controlador. Além do mais, há uma pessoa indicada pelo controlador, que faz a comunicação entre o controlador, titular – O titular dos dados é a pessoa física proprietária dos dados que serão tratados – e ANPD, para que todos tomem parte no processo de tratamento (BRASIL, 2018).

Após tomar conhecimento sobre os principais conceitos que compreendem a LGPD, é possível observar que essa lei forma uma metodologia para o tratamento dos dados, determinando o que pode ou não ser feito. Assim, a LGPD define alguns princípios, com o objetivo de evitar abusos no tratamento dos dados, que são (BRASIL, 2018):

- **Finalidade:** especificar e explicitar o objetivo do tratamento para os titulares dos dados;
- **Adequação:** tratar os dados adequadamente para a finalidade que foi solicitada ao titular;
- **Necessidade:** verificar se os dados associados a uma operação de tratamento são realmente pertinentes para a execução da tarefa;
- **Livre acesso:** garantir aos titulares o acesso detalhado às informações relacionadas às operações de tratamento;
- **Qualidade dos dados:** apresentar aos proprietários dos dados a exatidão, relevância e atualização dos dados associados a um processo de tratamento específico;

- **Transparência:** disponibilizar informações sobre o tratamento e os respectivos agentes de forma clara e de fácil acesso para o titular dos dados;
- **Segurança:** empregar medidas protetivas para salvaguardar os dados e evitar acesso indevido, vazamento ou perda;
- **Prevenção:** aplicar critérios para prevenir a ocorrência de eventos prejudiciais aos titulares dos dados. Aplicação de metodologias de segurança para melhorar os processos nas organizações, por exemplo, *International Organization for Standardization (ISO) 27000* ou *Control Objectives for Information and related Technology (COBIT)*;
- **Não discriminação:** evitar que as respostas aos tratamentos sejam ilícitas ou abusivas;
- **Responsabilização:** demonstrar que a organização está realizando o tratamento de forma adequada, e que há verificação contínua no ambiente interno para validar se as normas aplicadas são eficazes.

Por fim, para complementar as definições sobre a LGPD, vale ressaltar que, os direitos dos titulares apresentados pelo artigo 18 da lei, que são (BRASIL, 2018): (i) conhecer como os dados são recuperados; (ii) solicitar a exclusão dos dados, para que o tratamento seja finalizado; (iii) revogar os direitos que foram dados aos agentes de tratamento; (iv) retificar dados que foram tratados; (v) solicitar esclarecimento das decisões tomadas por algoritmos; (vi) os dados de uma plataforma podem ser portados para outra sem perdas; (vii) as decisões tomadas por meios automatizados podem ser revisadas. Portanto, com a integração da LGPD aos sistemas de saúde, bem como a blockchain, os pacientes estão assegurados por diversas garantias, assim, as organizações devem seguir uma série de regras para reduzir os riscos no tratamento dos dados dos pacientes.

2.3.2 Modelos para garantir privacidade

Para auxiliar na preservação da privacidade dos pacientes nos sistemas de saúde, pode ser aplicada a tecnologia blockchain provendo a pseudo-privacidade (FENG *et al.*, 2019). No entanto, essa característica não é suficiente para tornar os sistemas de saúde mais seguros e seguir a LGPD, uma vez que, a própria blockchain apresenta desafios sobre privacidade (BODKHE *et al.*, 2020). Alguns mecanismos combinados com a blockchain podem auxiliar a salvaguardar a privacidade dos pacientes nesses sistemas. São eles: (i) Criptografia Baseada em Atributos (CBA); (ii) Prova de Conhecimento Zero Não interativa (PCZNI); (iii) Criptografia homomórfica (CH); (iv) *Mixing* (MX); (v) K-Anonimato (KA); e (vi) Privacidade Diferencial (DF).

Criptografia baseada em atributos

A primeira técnica é a criptografia baseada em atributos, que apresenta uma solução para o desafio em compartilhar dados sensíveis utilizando políticas de acesso (GOYAL *et al.*, 2006).

Essa técnica emprega níveis de acesso para cada atributo criptografado e somente a categoria de usuários definidos na política de acesso poderá ler, escrever ou atualizar determinado atributo (GOYAL *et al.*, 2006). Para definir as políticas de criptografia de texto, são empregados quatro algoritmos básicos: configuração, codificação, decodificação e geração das chaves (WATERS, 2011). Portanto, o algoritmo da CBA pode ser descrito da seguinte forma (WATERS, 2011):

- **Configuração:** para o algoritmo de configuração é dado como parâmetro de entrada o conjunto de atributos e tem como retorno a chave mestra;
- **Codificação:** é o algoritmo para cifrar o texto, no qual é utilizado como parâmetro a chave pública, a mensagem e a estrutura de acesso. O retorno desse algoritmo é o texto cifrado, seguindo as políticas de acesso que foram definidas na estrutura de acesso;
- **Geração das chaves:** para gerar as chaves é empregada a chave mestra e o conjunto de atributos, assim, são geradas as chaves pública e privada;
- **Decodificação:** esse algoritmo é utilizado para decifrar o texto, que tem como parâmetro os atributos codificados e as chaves. Desse modo, os atributos são decifrados com base nas permissões definidas nas regras de acesso.

Contudo, essa técnica apresenta desafios que podem influenciar o custo ou a confiabilidade do sistema. Na CBA, o processo de revogação pode ser custoso, pois a atualização das políticas de acesso para cada atributo é uma tarefa complexa e pode não ocorrer por completo (AL-DAHMAN *et al.*, 2019). Quando aplicado a blockchain para saúde, pode ser um obstáculo, que os sistemas desse tipo prezam pela confiabilidade, nesse caso, se houver uma atualização de política incorreta, essa característica poderá ser violada.

Prova de conhecimento zero não interativa

Primeiramente, é importante entender o conceito de prova de conhecimento zero clássico, em que um agente tem o objetivo de provar que conhece a solução de um enigma sem apresentar a solução completa para outro agente (LI *et al.*, 2018). Um caso ilustrativo para compreender a definição de prova de conhecimento zero é o da caverna de Ali Baba, no qual temos dois agentes distintos, o provador Bob (P) e a verificadora Alice (V). O caso ocorre da seguinte forma, Bob conhece a chave secreta para abrir a porta no meio da caverna e quer provar para Alice que tem conhecimento da chave. Assim, Bob entra na caverna e Alice fica do lado de fora para que ele escolha um lado arbitrário (escolhendo o lado A) para entrar e sair pela porta que divide a caverna (no entanto, a caverna é circular). Desse modo Alice entra no início da caverna e grita o lado (escolhendo o lado B) em que Bob deverá sair. Portanto, Bob abre a porta e sai pelo lado escolhido por Alice (lado B). Por consequência, é possível provar que Bob conhece a chave secreta que abre a caverna, já que ele entrou e saiu por lados distintos dela (QUISQUATER *et al.*, 1990). Além disso, o conhecimento da chave é dado pela probabilidade de escolher um dos

lados, para, assim, o provador comprovar a hipótese de conhecimento da chave para o verificador (BELLARE; GOLDREICH, 1993).

No contexto do blockchain, em que há diversos pares que não se confiam, a prova de conhecimento zero convencional não é adequada, sendo válido utilizar a prova de conhecimento zero não interativa (FENG *et al.*, 2019). O esquema de Prova de Conhecimento Zero Não interativa (PCZNI) trabalha com uma prova unidirecional, em que somente o provador comunica com o verificador para apresentar a prova, a partir de uma *string* randômica. Se o provador e verificador tem uma *string* randômica comum, assim, é possível comprovar que o provador conhece o enigma, pois tem conhecimento do teorema gerador (BLUM; FELDMAN; MICALI, 1988). Em outras palavras, o provador disponibiliza a prova na rede e os verificadores distribuídos testam sua veracidade (BLUM; FELDMAN; MICALI, 1988). Os fundamentos da prova não interativa, com base em uma linguagem L, que pertence ao um conjunto não determinístico em tempo polinomial (NP), são (FENG *et al.*, 2019; BLUM; FELDMAN; MICALI, 1988):

- **Completeness:** representa a probabilidade de sucesso em provar que um teorema é verdadeiro, evidenciando o conhecimento sobre determinada informação é verdadeiro;
- **Solidez:** a probabilidade de êxito na prova de um teorema falso deve ser insignificante;
- **Conhecimento zero:** a prova não deve dar nenhuma informação completa, somente a veracidade do teorema.

Apesar de fornecer uma ótima alternativa para prover privacidade aos usuários dos sistemas de saúde, o PCZNI tem um alto custo computacional e isso pode ser um problema, quando integrado ao blockchain. A junção dessas duas técnicas, podem causar grandes despesas para os sistemas médicos, em virtude do custo de processamento, gerando alguns atrasos para enviar transações (FENG *et al.*, 2019; WU; WANG, 2014).

Criptografia homomórfica

Um esquema de criptografia que pode ser empregado ao blockchain, como discutem Feng *et al.* (2019), é a criptografia homomórfica, que funciona como um esquema de caixa preta. Com esse esquema é possível aplicar operações nas entradas e ainda fornecer informações úteis como resultado das operações. Por exemplo, considerando os salários de dois funcionários de uma empresa Z, sendo $F_1 = 1200$ e $F_2 = 1500$, cifrando os valores, há o seguinte resultado: $F_1 = 0,6$ e $F_2 = 0,75$. Aplicando a operação de tributação de 10% em cada salário, tem-se $F_1 = 0,54$ e $F_2 = 0,675$. Portanto, é possível observar que as operações seguem a mesma regra e possuem a mesma influência caso aplicadas aos salários reais. Em outras palavras, a unidade é diferente, mas tão representativa quanto a real.

A criptografia homomórfica aplicada a blockchain pode garantir a privacidade dos pares sem modificar duas principais características. As mensagens podem ser criptografadas em cada

ponta na troca dos ativos e somente ser divulgado publicamente o resultado das operações realizadas entre os usuários, sem comprometer os dados reais (ZHANG; XUE; LIU, 2019). Além disso, essa técnica é recente e vêm sendo estudada e evoluída, assim, apresenta alguns desafios, como a realização de diversas operações simultâneas. Aplicar essas operações seria inviável, pois uma poderia cancelar a outra, tornando o sistema inseguro (WILL; KO, 2015). Para o caso do blockchain para saúde, seria uma limitação, haja vista que esses sistemas necessitam de alta confiabilidade e segurança (FENG *et al.*, 2019).

Mixing

A técnica de *Mixing* está associada à preservação da privacidade da identidade dos usuários, com base na randomização do caminho de origem durante a troca de mensagens (FENG *et al.*, 2019). Essa técnica utiliza-se de uma caixa decisora entre o remetente e o receptor, com o propósito de randomizar o caminho de origem da mensagem, para que usuários maliciosos se confundam com a origem, sendo difícil descobrir a identidade do remetente (FENG *et al.*, 2019; ZHANG; XUE; LIU, 2019). Tal abordagem pode ser válida para anonimizar as identidades na blockchain, tendo em vista que, as carteiras digitais são pseudo-anônimas e assim é possível inferir o usuário que enviou uma mensagem (ZHANG; XUE; LIU, 2019). O *Mixing* apresenta duas estruturas, a centralizada e descentralizada, no qual a descentralizada pode contribuir com o blockchain, eliminando o ponto único de falha (FENG *et al.*, 2019). Algumas ferramentas baseadas em blockchain que empregam essa técnica são o CoinJoin e Mixcoin, que têm o objetivo específico de realizar operações de pagamento na rede do Bitcoin (ZHANG; XUE; LIU, 2019). No entanto, esta também apresenta limitações, como a ligação de um atributo público como valor da moeda ao usuário que o enviou, assim, o *Mixing* não trata ataques de ligação aos registros, para inferir o usuário real (RUFFING; MORENO-SANCHEZ, 2017).

K-Anonimato

O K-Anonimato é uma técnica útil para anonimizar os dados e evitar a re-identificação de um registro a partir de outro público, gerando a incerteza no conhecimento completo do dado (BRITO; MACHADO, 2017; SWEENEY, 2002). Essa técnica pode proteger as bases de dados contra os ataques de ligação e os valores de K indicam o nível de privacidade inserido nos dados. Além disso, quanto maior for o K , mais aleatória a base é, isso implica na utilidade dos dados, que são inversamente proporcional aos valores de K , isto é, dados muito privados podem ser pouco úteis (SWEENEY, 2002; CHEN *et al.*, 2009).

Para aplicar a anonimidade existem alguns métodos, como o da generalização e supressão (CHEN *et al.*, 2009). A generalização transforma o dado original em classes que podem representar o dado um nível acima, por exemplo: considerando um paciente com 25 anos pode ser generalizado como uma classe de idade entre 20 e 30 anos. Já a supressão, parte da *string* que representa o dado, que é substituído por um símbolo, como “*”, com o $K=2$ e um CEP arbitrário 86350 é modificado para 863** (CHEN *et al.*, 2009).

Além disso, esse método pode ser útil para aplicar ao blockchain, pois pode mitigar os ataques de ligação submetidos aos registros, e combinado a uma aplicação de saúde, pode prover maior privacidade (WANG *et al.*, 2018). Contudo, há desafios na aplicação desse método, uma vez que não pode prevenir ataques de ligação submetidos a atributos relacionado a um usuário específico ou ataques probabilísticos (BRITO; MACHADO, 2017; BARBOSA; BRITO; ALMEIDA, 2020). Esses São ataques que definem a probabilidade de uma mensagem enviada no sistema pertencer a um usuário, mesmo que seja privado, para assim descobrir o usuário real que a enviou (BAGAI; TANG; KIM, 2013).

O K-Anonimato também pode apresentar vulnerabilidades envolvendo ataques, uma vez que, é sensível ao ataque de homogeneidade e ao ataque de conhecimento de segundo plano (MACHANAVAJJHALA *et al.*, 2007). O ataque de homogeneidade aproveita dos grupos de registros que tenham atributos em comum, para descobrir com exatidão o significado de um atributo pertencente a um usuário (MACHANAVAJJHALA *et al.*, 2007). Por exemplo, Bob apresentou problemas de saúde e foi hospitalizado no dia 20 de dezembro, Alice, sua vizinha, por curiosidade gostaria de saber qual o problema de saúde de Bob. Desse modo, Alice consulta os registros anonimizados que foram disponibilizados pelo hospital no dia 20 de dezembro. Ela conhece Bob e sabe que ele é um homem de 31 anos e reside no CEP 13566. Com base nessas informações, Alice descobre que três dos registros tem os atributos $CEP = 135 **$, $idade = \text{entre } 30 \text{ e } 40$, $sexo = M$ e $doença = \text{diabetes}$, logo, Bob tem diabetes (MACHANAVAJJHALA *et al.*, 2007). O ataque de conhecimento de segundo plano se baseia em um conhecimento anterior de fatos, que auxiliam na identificação de atributos referentes a um indivíduo (MACHANAVAJJHALA *et al.*, 2007). Por exemplo, Alice tem uma amiga, a Eve, que foi hospitalizada recentemente no mesmo hospital que Bob, ela sabe Eve é mulher de 23 anos e reside no CEP 13568. Assim, Alice encontra uma base anonimizada dos pacientes hospitalizados no mesmo dia que Eve e, descobre que quatro registros contém os seguintes atributos: $CEP = 135 **$, $idade = \text{entre } 20 \text{ e } 30$ e $sexo = F$. O atributo que descreve a doença, em dois deles representam infecção viral e nos outros dois doença cardíaca, no entrando, se sabe que a incidência de problemas cardíacos em jovens é baixa, logo, Eve pode ter uma infecção viral (MACHANAVAJJHALA *et al.*, 2007).

Privacidade Diferencial

A Privacidade Diferencial (DP) é outro método que pode ser aplicado para anonimizar dados, e contribuir para esta finalidade quando aplicado a blockchain. Essa técnica também aplica ruído aos valores retornados com base em uma consulta aos registros dos usuários (LU; AU, 2017). Para ilustrar o modelo de DP, pode-se considerar a seguinte analogia, seguindo duas bases de dados médicas: a primeira com registros médicos pessoais de um paciente e a segunda com registros financeiros deste mesmo paciente. Desse modo, poderia ser associada uma tabela pública de saúde com outro banco de dados que contém o número do cartão de crédito do paciente, para, assim, fazer um ataque de ligação para descobrir tais dados e fraudar o cartão

(DWORK; ROTH, 2014; LU; AU, 2017).

Diferentemente do K-Anonimato, a Privacidade Diferencial consegue prover a privacidade para contra ataques de ligação à tabela e probabilísticos (BARBOSA; BRITO; ALMEIDA, 2020). Assim, essa característica trabalha com dados estatísticos, ou seja, a maioria das propostas realizam operações somente em dados agregados, com alta granularidade (DWORK; ROTH, 2014). Por exemplo, trabalhos que empregam esta técnica utilizam-se de bases estatísticas somente para retornar uma consulta específica, como média, mediana, desvio padrão e outras (CLIFTON; ANANDAN, 2013). Portanto, os dados retornados dependendo da finalidade não são tão úteis para o usuário final, mas têm alta privacidade.

A definição formal da Privacidade Diferencial descreve como essa técnica funciona, calculando a diferença entre duas consultas em bases de dados similares, que diferem em um elemento (DWORK; ROTH, 2014). Assim, a função aleatória M (mecanismo de privacidade) é ϵ -*differential-privacy*, se dois conjuntos de dados B_1 e B_2 diferem em no máximo um elemento em uma saída S (DWORK; ROTH, 2014). A Equação 2.1 apresenta o formalismo para representar a privacidade diferencial, em que o parâmetro ϵ **representa o nível de ruído** inserido no dado, que quanto menor o ϵ mais ruído é inserido (LU; AU, 2017). O valor S é uma saída e M é um mecanismo de privacidade (LU; AU, 2017).

$$Pr[M(B_1) \in S] \leq \exp(\epsilon) \times Pr[M(B_2) \in S]. \quad (2.1)$$

O mecanismo M é um modelo probabilístico que tem como base uma distribuição densidade de probabilidade para randomizar os dados e deixá-los indistinguíveis, partindo de dois mecanismos principais: Laplaciano e Exponencial (DWORK; ROTH, 2014). A Equação 2.2, denota o mecanismo Laplaciano, cujo valor x é a entrada da distribuição e b é uma constante que pode ser ajustada pelo criador do modelo.

$$Lap(x|b) = \frac{1}{2b} e^{-\frac{|x|}{b}}. \quad (2.2)$$

Ademais, para representar o quanto a base é sensível a ataques, é calculada a sensibilidade, que resulta em um índice que tem influência no ruído inserido sobre a consulta (DWORK; ROTH, 2014). Formalmente, a sensibilidade é representada por Δf e pela Equação 2.3. O resultado da sensibilidade é aplicado para calcular a constante b , utilizada na Equação 2.2, para o mecanismo Laplaciano, a partir da fórmula $b = \frac{\Delta f}{\epsilon}$ (DWORK; ROTH, 2014). Por fim, para aplicar o ruído à consulta, baseado na adição de ruído segue a fórmula: $f_{ruído}(d) = f(d) + Lap(\frac{\Delta f}{\epsilon})$.

$$\Delta f = \max_{d_1, d_2} (|f(d_1) - f(d_2)|) \quad (2.3)$$

METODOLOGIA E TÉCNICAS DE PESQUISA

Este capítulo tem o propósito de apresentar como a pesquisa foi estruturada de forma sistemática, seguindo os conceitos discutidos em [Gil \(2002\)](#) para uma pesquisa experimental. Primeiramente, foi realizado um levantamento na literatura para encontrar trabalhos semelhantes, bem como os principais conceitos que envolvem o tema pesquisado, que são blockchain, privacidade, saúde e outros. Em seguida foram definidas as questões de pesquisa com base nos achados da literatura, a fim de selecionar os métodos utilizados para definir o tipo de rede, protocolo de consenso e a estrutura DLT. Para descrever como foi conduzido o processo de pesquisa, assim como a metodologia empregada, discutida em [Gil \(2002\)](#), foram seguidas as seguintes tarefas:

1. Levantamento bibliográfico;
2. Formulação das questões de pesquisa;
3. Seleção da estrutura de livro-razão distribuído para a proposta;
4. Seleção do modelo de rede blockchain;
5. Levantamento do protocolo de consenso e plataformas;
6. Definição das métricas para execução experimental.

3.1 Levantamento bibliográfico

A etapa de levantamento bibliográfico tem o intuito de explorar a literatura sobre o tema blockchain aplicado a saúde, além de compreender a preservação da privacidade nesse ambiente. Primeiramente, é elaborada a conscientização do problema, com objetivo de levantar o máximo de informações que possam contribuir para definição do problema. Assim, foram buscados os conceitos associados ao tema e abordagens empregadas. Também foi realizada a análise da

viabilidade da problemática da pesquisa, explorando as possibilidades na literatura e examinado os aspectos técnicos relacionados ao blockchain, tal como a tecnologia na saúde, para que fosse possível adquirir um conhecimento preliminar sobre o escopo geral da pesquisa.

Desse modo, para esta pesquisa foram levantados os estudos e conceitos relacionados utilizando o motor de busca *Engineer Village*¹, que indexa as bases bibliográficas da *IEEE*, *Springer*, *ACM* e *Elsevier*. Para complementar, a pesquisa foi realizada mais uma vez utilizando dos motores de busca e bases de dados bibliográficas *IEEE Xplore*, *ACM digital Library*, *Science Direct* e *PubMed*. Com os repositórios de artigos definidos, o passo seguinte foi definir algumas palavras-chave para conduzir o processo de busca, como é apresentado na Tabela 1.

Tabela 1 – Número de trabalhos associados com as palavras-chave utilizadas na revisão bibliográfica, indicando uma sequência de estudos a partir das palavras-chave gerais para as específicas.

Ano	Palavras-chave					
	blockchain	"private blockchain"	blockchain and health*	blockchain and privacy	blockchain and privacy and health*	blockchain and imag*
2015	41	0	1	5	0	5
2016	171	1	11	43	3	5
2017	670	6	52	190	22	19
2018	2056	32	150	407	48	72
2019	4710	16	99	231	43	57
2020	5422	42	269	571	99	96
Total	13070	97	571	1257	215	254

A quantidade de trabalhos encontradas por palavras-chave, são apresentados na Tabela 1, que seguiram os seguintes critérios de inclusão: artigos somente em inglês e publicações dos últimos 5 anos (2015 até 2020). O levantamento dos trabalhos ocorreu semanalmente, dessa forma, os motores de busca foram configurados para enviarem toda a sexta-feira as atualizações. Com base na exploração na literatura, foi possível definir e responder às questões de pesquisa secundárias e assim contribuir para resolução da questão primária.

3.2 Formulação das questões de pesquisa

Após a exploração do tema blockchain em ambientes de saúde, como o intuito de prover privacidade, foi realizada a formulação das questões de pesquisa, que tem o objetivo de buscar e identificar as limitações dos trabalhos semelhantes na literatura. O foco das questões de pesquisa é compartilhar dados de saúde preservando a privacidade dos pacientes que os disponibilizam, como também responsabilizar as organizações pelos possíveis vazamentos de dados. Desse modo, as questões de pesquisa levantadas foram divididas em questão principal (QP) e questões secundárias (QS), que complementam a primária. São elas:

- **QP.** Como mitigar os riscos à privacidade dos pacientes no compartilhamento de dados de saúde em redes baseadas em blockchain?

¹ www.engineeringvillage.com

- **QS 1.** Como rastrear ativos compartilhados, com o objetivo de responsabilizar os agentes de tratamento pelo uso inadequado dos dados?
- **QS 2.** Como mitigar ataques de re-identificação para prover privacidade aos pacientes em ambientes de saúde, quando estes compartilham seus dados?
- **QS 3.** Qual a estrutura de dados, baseada na tecnologia de livro-razão distribuído é adequada para esta proposta?
- **QS 4.** Qual modelo de rede blockchain pode proporcionar o melhor custo benefício para aplicações na área da saúde?
- **QS 5.** Qual modelo de privacidade pode contribuir para melhorar a privacidade em blockchain?

Além disso, as questões de pesquisa tiveram como base alguns casos de ataques aos ambientes de saúde, bem como violações, com o intuito de fraudar ou vazar dados. Alguns desses ataques mais recentes foram: (i) o vazamento de 68 mil registros médicos no hospital metodista em Indiana nos Estados Unidos, em outubro de 2019, e (ii) o vazamento de cerca de 50 mil de registros de pacientes no Minnesota hospital, no início de 2020². Cada uma das questões de pesquisa se complementam com o propósito de responder à questão principal e atingir os objetivos deste trabalho, que é desenvolver um *framework* para compartilhar dados de saúde e reduzir riscos a privacidade dos usuários.

3.3 Seleção da estrutura de dados a partir da tecnologia de livro-razão distribuído

A tecnologia de livro-razão distribuído (do inglês, DLT) apresenta algumas estruturas para organizar os dados. Para esta proposta foram selecionados duas dessas estruturas para serem analisadas: o blockchain (NAKAMOTO, 2008) e o Tangle (POPOV, 2018). O blockchain foi selecionado por ser uma das estruturas mais utilizadas na literatura, apresentando o potencial de ser integrado em diferentes aplicações, tal a área da saúde, como discutem os seguintes autores Radanović e Likić (2018), Dabbagh, Sookhak e Safa (2019) e McGhin *et al.* (2019). O Tangle foi escolhido como fator de comparação com o blockchain, pois este promete o processo de consenso mais rápido, devido ao fato de possuir um protocolo de consenso probabilístico (POPOV, 2018). Isso porque em sistemas de saúde o processo de consenso ágil pode contribuir na velocidade da resposta, por serem sistemas críticos e necessitarem de uma tomada de decisão rápida.

Com base nas duas estruturas de dados, o Tangle (baseado em grafos) e o blockchain (baseado em blocos encadeados), foram analisados para decidir aquela que melhor se adaptava a

² <https://portswigger.net/daily-swig/data-breach-at-minnesota-hospital-threatens-nearly-50-000-healthcare-records>

esta proposta. Após o levantamento das principais vantagens das estruturas, foi selecionada a blockchain, uma vez que é a mais consolidada na literatura e na indústria, com casos de sucesso em aplicações de saúde (AZARIA *et al.*, 2016; PATEL, 2018). Além disso, o blockchain tem suporte a algumas plataformas para auxiliarem na implementação, como o Hyperledger Fabric e o Ethereum. Já o Tangle apresenta menor confiabilidade, pois ainda não está consolidado e não apresenta plataformas que auxiliem na implementação, tornando sua implementação menos confiável em ambientes de saúde que tratam de dados sensíveis. Em suma, esta fase auxilia na resposta da questão **QS 3** (Qual a estrutura de dados, baseada na tecnologia de livro-razão distribuído é adequada para esta proposta?), isto é, a estrutura adequada para esta proposta é o blockchain.

3.4 Seleção do modelo de rede blockchain

Com a escolha da estrutura de dados, a blockchain, foi necessário selecionar os modelos de redes, seguindo os que foram discutidos na Subseção 2.1.3. Entre essas redes, os modelos podem ser públicos e privados, além disso, os mecanismos de consenso podem ser permissionados e não-permissionados (CORTEN, 2018). Para decidir qual o modelo de rede blockchain era o mais adequado para esta proposta, foi aplicado o fluxograma apresentado na Figura 3, que se baseia nas pesquisas de Lo *et al.* (2017) e Wust e Gervais (2018).

Vale ressaltar que, para o fluxograma da Figura 3 foram destacados os atores: profissional da saúde, paciente, pesquisador e provedor de saúde. Em vista disso, os conflitos de interesse estão associados aos pacientes que compartilham seus dados, bem como os profissionais de saúde, pesquisadores e provedores de saúde os acessam. Assim, os dados quando recuperados sem autorização por terceiros (pesquisadores e profissionais da saúde) podem causar vazamentos, indisponibilidade e adulteração. Nesse contexto, o fluxograma tem o objetivo de apresentar a motivação da utilização do blockchain e qual processo de consenso é o mais adequado para esta proposta.

Com o objetivo de validar e analisar a utilização da blockchain em aplicações que compartilham dados de saúde, com base no fluxograma (Figura 3), os seguintes tópicos foram respondidos:

1. **Os dados precisam ser armazenados?** Sim. O *framework* pretende compartilhar dados entre várias organizações de saúde, logo, todos os dados necessitam ser armazenados. Por exemplo, os *logs* e metadados das imagens compartilhadas necessitam ser armazenadas para os usuários da rede acessarem;
2. **Permite múltiplos escritores?** Sim. O objetivo do *framework* é o compartilhamento e acesso aos dados de saúde de forma segura, sendo assim, diversos pares podem escrever

Figura 3 – Fluxograma para a escolha do tipo de rede blockchain e processo de consenso. Adaptado de Wust e Gervais (2018); Lo *et al.* (2017).

no livro-razão de forma descentralizada. Por exemplo, diversos nós como provedores de saúde e pesquisadores podem escrever no livro-razão;

3. **Necessita somente de operações centralizadas?** Não. Os pares da rede estão descentralizados e distribuídos fisicamente e as operações são realizadas dessa forma. Por exemplo, há diversas bases de dados distribuídas entre os centros de pesquisa e hospitais. As operações são realizadas por cada organização e podem ser compartilhadas entre diversos usuários, um modelo descentralizado seria conveniente para essas operações;
4. **Os pares que modificam a rede são conhecidos?** Sim. Os pares que executam a tarefa de validação são organizações de saúde que tratam de dados sensíveis, dessa forma, compõem um conjunto confiável para a rede. Se pares desconhecidos pudessem participar da validação, seria arriscado, uma vez que, a rede estaria mais suscetível a usuários maliciosos;
5. **Os pares que escrevem na rede são confiáveis?** Não. Se os pares da rede fossem totalmente confiáveis e controlados, não haveria necessidade da blockchain. Consequentemente, em sistemas de saúde, não é possível afirmar que todos os pares que compõem a rede são confiáveis, tendo em vista que são heterogêneos. O processo de gerenciá-los e torná-los confiáveis é complexo, logo, a blockchain pode contribuir para mitigar esse problema. Os nós são concorrentes, pois desconfiam mutuamente de si, um pesquisador pode ter sua máquina infectada e tornar um nó malicioso, tentando invalidar as transações;

6. **A auditoria dos dados é pública?** Não. Por se tratar de sistemas de saúde, a publicação indevida de dados pode ocasionar fraudes ou vazamentos. Os dados gerenciados são de pacientes (dados pessoais sensíveis) e necessitam de um modelo privado. Somente um conjunto específico de organizações podem acessar os dados de forma pública.

Seguindo o fluxograma da Figura 3, complementando com a questão **QS 4** (Qual modelo de rede blockchain pode proporcionar o melhor custo benefício para aplicações na área da saúde?), esclarecida nos tópicos anteriores, pode-se afirmar que o modelo de rede mais adequado é a blockchain **permissionado privado**. Esse modelo oferece maior controle sobre os pares, que podem estabelecer um grupo confiável de validadores, sem que os dados sejam disponibilizados para todos os usuários acessarem. Vale destacar que, a justificativa da utilização da blockchain para o cenário de saúde deste trabalho é a rastreabilidade. Com a blockchain todos os dados acessados podem ser rastreados e tornam-se imutáveis, além de mitigar riscos a adulteração dos *logs*.

3.5 Seleção do protocolo de consenso e plataforma

Com o modelo da blockchain definido, o próximo passo foi determinar o protocolo de consenso mais adequado para este trabalho, com o propósito de complementar a resposta da questão **QS 4**. Dessa forma, o protocolo de consenso deve seguir o modelo permissionado privado, que provê os benefícios descritos pelo fluxograma da Figura 3. Portanto, foi selecionando o protocolo de consenso PBFT, que garante um modelo privado de acordo entre os pares, e que tem como validadores um conjunto confiável de pares pré-estabelecidos (ANDROULAKI *et al.*, 2018). Com esse protocolo, os sistemas de saúde podem ser mais confiáveis, sem ocasionar grandes custos a rede para validar um bloco, como é o caso do Bitcoin (LIANG *et al.*, 2017; ICHIKAWA; KASHIYAMA; UENO, 2017).

Para facilitar a implementação e proporcionar maior confiabilidade na construção da rede, com base em uma plataforma estabelecida, foi utilizado o Hyperledger Fabric (HLF) da IBM e Linux Foundation. Essa plataforma suporta o protocolo de consenso PBFT e apresenta uma rede com características privadas, além de ser customizável como os contratos inteligentes (ANDROULAKI *et al.*, 2018). O HLF aplica um protocolo de validação com baixo custo, por trabalhar com uma abordagem de votação (CASTRO; LISKOV, 1999; ANDROULAKI *et al.*, 2018). Além disso, ele também fornece ferramentas para construção de contratos inteligentes em diferentes linguagens – a esse respeito, neste trabalho foi utilizado a linguagem Go. Em suma, o HLF proporciona facilidades para construção e integração da blockchain para o ambiente de saúde, tal como para um modelo de rede permissionada privada.

Por fim, esta etapa auxilia na resposta da **QS 4** (Qual modelo de rede blockchain pode proporcionar o melhor custo benefício para aplicações na área da saúde?), que tem como resposta

o modelo **permissionado privado**, com o mecanismo de consenso PBFT, além da plataforma HLF, para facilitar a implementação da rede.

3.6 Método de avaliação experimental

Esta seção tem o objetivo de descrever os métodos empregados para conduzir e executar os experimentos desta pesquisa. Os componentes considerados nas simulações sobre a rede blockchain tem o propósito de verificar o desempenho e suas influências sobre a aplicação deste *framework*. Os modelos de privacidade analisados também foram considerados no que se refere ao custo benefício de cada método, aplicando avaliações teóricas e quantitativas. Esta etapa auxilia na resolução das questões **QS 2** (Como mitigar ataques de re-identificação para prover privacidade aos pacientes em ambientes de saúde, quando compartilham seus dados?) e **QS 5** (Qual modelo de privacidade pode contribuir para melhorar a privacidade em blockchain?), em que é definido o modelo de privacidade, bem como o método para avaliá-los.

Avaliação da rede blockchain

A primeira etapa é a avaliação experimental da blockchain, com o intuito de analisar a carga gerada na rede, assim como o consumo das máquinas que o executam. As métricas adotadas para desempenho da rede foram: (i) latência; (ii) vazão (*throughput*); e transações por segundo (TPS) (SADIKU; MUSA, 2013; SHAHNAZ; QAMAR; KHALID, 2019). As métricas de desempenho do hardware das máquinas foram: (i) uso de CPU; e (ii) uso de memória (SADIKU; MUSA, 2013; SHAHNAZ; QAMAR; KHALID, 2019). Para a blockchain foi empregado o Hyperledger Fabric, logo, pode ser utilizada a ferramenta Hyperledger Caliper³ para facilitar o processo de simulação. Assim, essa ferramenta fornece de forma facilitada suporte as métricas citadas e é possível configurar o número de usuários, quantidade de replicações dos experimentos, processo de consenso e outras tarefas (SANKAR; SINDHU; SETHUMADHAVAN, 2017).

Avaliação dos modelos de privacidade

Na avaliação dos modelos de privacidade foram considerados aspectos quantitativos e qualitativos, estes sendo respectivamente associados a métricas para validar a condição do sistema, o nível de privacidade e a utilidade dos dados privados. Também foi analisado como a proposta impacta e auxilia na mitigação de problemas de privacidade em sistemas de saúde, a fim de responder à questão **QS 2**. Com o propósito de responder essa questão, foram selecionados os modelos de anonimização K-Anonimato e Privacidade Diferencial, para, assim, prover privacidade aos pacientes.

Na avaliação qualitativa foi levantada uma lista de requisitos de privacidade que o sistema necessita cobrir para prover melhor privacidade aos usuários. Para esse levantamento foi empregado o projeto da *Open Web Application Security Project* (OWASP) Foundation, que

³ <https://hyperledger.github.io/caliper/>

elencar os 10 principais riscos à privacidade para aplicações na Internet. Os riscos apresentados pela OWASP fazem paralelo com os riscos enfrentados pelas organizações de saúde e esta proposta tem o objetivo de mitigá-los (OWASP, 2016). Além disso, em paralelo com esses riscos foi integrado o conceito de *Privacy by Design* (PbD) que visa modelar projetos de sistemas para prover privacidade aos seus usuários, em que foi empregado o *framework* com o propósito de responder parte da questão de pesquisa **QS 5** (BARBOSA; BRITO; ALMEIDA, 2020; CAVOUKIAN, 2009).

Para avaliar os métodos de privacidade, foram abordadas as métricas da literatura que fossem adequadas para validar os modelos K-Anonimato e Privacidade Diferencial, além de auxiliar na resolução da questão de pesquisa. A partir do trabalho de (WAGNER; ECKHOFF, 2018) foram utilizadas as métricas de entropia e utilidade – no qual a utilidade é representada pela distância entre o atributo original e o anonimizado. A métrica de entropia (veja a Equação 3.1) é análoga ao nível de privacidade, pois indica a desordem dos atributos anonimizados, bem como a probabilidade de um nó malicioso inferi-lo, com base em um conhecimento prévio sobre os dados (WAGNER; ECKHOFF, 2018). A métrica de utilidade é representada pela distância de cossenos (ver Equação 3.2), uma vez que, é distância que melhor representa atributos categóricos e tem o objetivo de quantificar quanto um dado é útil após anonimizado. Portanto, com a avaliação quantitativa é possível validar a resposta da questão de pesquisa **QS 5** (Qual modelo de privacidade pode contribuir para melhorar a privacidade em blockchain?) e quantificá-la numericamente. Por fim, para melhor descrever as equações, na Equação 3.1 temos o $p(x)$ que é a probabilidade de cada membro do conjunto de anonimato x . Na Equação 3.2, u é o vetor que corresponde aos dados originais, sem anonimizar, v é o vetor que representa os dados anonimizados e $\|u\|_2 \cdot \|v\|_2$ é a norma desses vetores.

$$privacidade \equiv H(x) = - \sum_{x \in X} p(x) \log_2 p(x). \quad (3.1)$$

$$dist(u, v) = \frac{u \cdot v}{\|u\|_2 \cdot \|v\|_2}. \quad (3.2)$$

$$utilidade \equiv dissimilaridade = 1 - dist(u, v).$$

Vale ressaltar que, as métricas representadas pelas Equações 3.1 e 3.2 descrevem a incerteza do dado após anonimizado e a dissimilaridade. A incerteza auxilia na análise do dado após anonimizado, verificando se ainda é possível extrair alguma informação dos registros para contribuir na identificação do indivíduo (WAGNER; ECKHOFF, 2018). Além disso, a métrica de utilidade retrata a dissimilaridade dos dados, considerando a equação da dissimilaridade entre o dado anonimizado e o não anonimizado, isto implica em uma métrica para quantificar o quanto um dado é útil após anonimizado, para, assim, complementar a entropia (WAGNER;

ECKHOFF, 2018; BERTINO; FOVINO; PROVENZA, 2005). Essas métricas são inversamente proporcionais, uma vez que, quanto maior a entropia maior privacidade o dado terá, por sua vez, a utilidade dele será menor. Por exemplo, um dado criptografado tem alta entropia, visto que o objetivo é que não seja interpretado por um usuário malicioso.

3.7 Considerações finais

Este capítulo apresentou as técnicas levantadas para a condução desta pesquisa. Estabelecendo o método para a coleta dos estudos relacionados, a delimitação do escopo e a formulação das questões de pesquisa. Com base nas questões formuladas, foram buscados métodos para respondê-las, a fim de definir quais os modelos blockchain seriam utilizados, qual o processo de consenso e avaliação experimental. Desse modo, este capítulo foi importante para responder parte das questões de pesquisa fundamentada na literatura e enumerar as métricas que trabalhos correlatos utilizam e são úteis para validar a proposta, considerando os aspectos da blockchain e da privacidade.

TRABALHOS RELACIONADOS

Este capítulo tem o propósito de investigar a literatura e discutir a respeito do tema blockchain para ambientes de saúde. Assim, foram coletados trabalhos com dois focos específicos: o compartilhamento de informações de saúde, e a proteção da privacidade nos ambientes de saúde baseados em blockchain. A última etapa deste capítulo tem o objetivo de enumerar os trabalhos mais relevantes, de modo a compará-los com esta proposta, listando as semelhanças, diferenças e os pontos de contribuição desta pesquisa. A investigação dos trabalhos relacionados, é uma das fases da metodologia desta pesquisa que, foi fundamentada em bases de dados e motores de busca conceituados (*IEEE Xplore*, *ACM Digital Library* e outras). Também vale ressaltar que, para o levantamento bibliográfico, foi tomado como parâmetro alguns trabalhos de revisão sistemática como [Radanović e Likić \(2018\)](#), [Hölbl *et al.* \(2018\)](#) e [Shi *et al.* \(2020\)](#). Diante da pesquisa realizada nos repositórios de artigos, foram selecionados os principais trabalhos associados ao tema compartilhamento de informações de saúde baseado em blockchain, bem como sobre privacidade nesse cenário. Por fim, para seleção dos trabalhos os critérios adotados foram artigos somente de periódicos. Estes trabalhos serão descritos a seguir.

4.1 Compartilhamento de informações de saúde baseado em blockchain

O primeiro trabalho que será discutido é o *MedRec: Using Blockchain for Medical Data Access and Permission Management*, sendo um dos mais citados sobre o assunto e proposto por [Azaria *et al.* \(2016\)](#). O MedRec é uma estrutura baseada em blockchain para armazenar registros médicos eletrônicos. O trabalho tem a finalidade de resolver alguns problemas como o acesso lento aos dados, interoperabilidade e uma melhor qualidade para a disponibilização dos dados em pesquisas médicas. [Azaria *et al.* \(2016\)](#) desenvolvem a proposta com base em uma rede P2P privada, bem como integra contratos inteligentes por meio da plataforma Ethereum, a partir do consenso não permissionado. Assim, é possível gerenciar e rastrear as transições de estados

dos ativos na rede. Essa abordagem fornece ao paciente uma agência para consulta de todo seu histórico médico, deixando ele informado sobre as decisões médicas. Os dados disponibilizados são flexíveis para serem moldados a fim de seguir padrões de dados abertos. Em suma, o MedRec é uma abordagem viável para a aplicação em sistemas de gerenciamento de dados de saúde, proporcionando maior segurança e uma linguagem comum para o compartilhamento dos dados para pesquisa. Os autores têm como próximos passos realizarem testes e a análise com uma gama de usuários diversificados. Além disso, uma preocupação da proposta é sobre a privacidade dos pacientes, uma vez que, os autores não consideram fortemente esse aspecto.

O trabalho de [Xia et al. \(2017\)](#) *MeDShare: Trust-Less Medical Data Sharing Among Cloud Service Providers via Blockchain* propõe uma abordagem para o compartilhamento de registros de saúde, com o objetivo de garantir a segurança dos dados utilizados em análise de *big data*. O MeDShare apresenta características como controle de acesso, verificação da confiabilidade dos pares da rede e identificação de usuários maliciosos com base na blockchain. Para modelar o sistema proposto pelos autores, foram definidas algumas camadas: (i) **camada do usuário**: define a característica do usuário que irá acessar o sistema; (ii) **camada de consulta dos dados**: define estruturas de consultas para acessar os dados; (iii) **estrutura de dados e camada de procedência**: auxilia na revocação de acesso e disponibilização das informações, com um banco de dados seguro e um mecanismo que apresenta a origem dos dados, a fim de fornecer procedência; (iv) **infraestrutura de banco de dados**: faz o gerenciamento da organização dos dados os associando a blockchain para armazenar informações. Consequentemente, os autores projetam um sistema intermediário ao provedor de nuvem, que auxilia o controle de acesso e auditabilidade dos dados com a blockchain. Para a implementação da proposta foi empregada a plataforma Ethereum, cujo contrato foi construído na linguagem Solidity. Desse modo, para mensurar o desempenho da proposta foi empregado o *plugin* JMeter, além da fundamentação em um estudo de caso real para a transferência de registros médicos entre provedores de nuvem. Assim, o MeDShare resulta em uma alta latência na simulação, com um número baixo de usuários implicando em atrasos para resposta devido à carga gerada na rede. Ademais, com o processo de revocação de acesso, os autores apresentam vulnerabilidades, que podem deixar os sistemas médicos sob risco.

Ainda sobre o compartilhamento de registros médicos, em [Shen, Guo e Yang \(2019\)](#), o *MedChain: Efficient Healthcare Data Sharing via Blockchain* aborda um *framework* flexível e escalável para a troca de dados médicos para estudo. Os principais diferenciais desse trabalho em relação à literatura da área, são: (i) uma estrutura flexível para o compartilhamento de dados, que seja segura e possibilite a troca de diferentes categorias de dados de pesquisa; (ii) uma abordagem para verificação das integridades das chaves utilizadas no compartilhamento de dados para dispositivos IoT para saúde; e (iii) melhor eficiência na auditabilidade, integridade e segurança dos registros compartilhados. Os autores desenvolvem uma estrutura composta por dois pares, os superpares e os pares de borda, que respectivamente são os provedores de saúde e os usuários de saúde menores. Essa estrutura fornece um esquema de gerenciamento

de chaves, auxiliado pelo blockchain com o protocolo de consenso BFT-SMaRt (SOUSA; BESSANI, 2012), para realizar o compartilhamento dos dados. Para provar a eficiência da proposta, os autores realizam experimentos relacionados ao desempenho da blockchain e à segurança que a estrutura garante. Os resultados para avaliação da segurança, seguindo uma análise teórica provam que a estrutura garante aos usuários de saúde integridade dos dados, resistência a ataques de replicação, resistência a ataques de máscara e não repúdio; e proteção da privacidade por encriptação da informação. Com relação aos resultados de desempenho, apresenta-se um gasto aceitável, contudo, é maior que os custos apresentados pelos sistemas sem blockchain. As limitações apresentadas pelo trabalho são: para transferir os dados, é necessário realizar ações dos provedores de saúde, que possuem os dados dos pacientes; as informações são inseridas manualmente para ficarem disponíveis para transferência; e a privacidade dos usuários é protegida somente por meio de chaves criptográficas.

Em Patel (2018) é abordado o trabalho *A framework for secure and decentralized sharing of medical imaging data via blockchain consensus* que explora o compartilhamento de imagens médicas usando a blockchain. O autor propõe uma estrutura para que os pacientes possam compartilhar suas imagens médicas de forma segura e controlada. O esquema para construção da rede segue a estrutura *Image Share Network (ISN)*, da Rede Nacional de Radiologia Norte americana, que serve como apoio para construir a proposta, com base nos problemas apresentados por essa rede centralizada de compartilhamento. A ideia do trabalho é registrar uma lista de estudos, bem como uma lista dos pacientes aos quais esses estudos pertencem. Assim, o proprietário do estudo (paciente) poderá selecionar com quem deseja compartilhar seus dados. No processo de compartilhamento de imagens são utilizados conceitos de criptografia de chave pública para protegê-las, além da blockchain utilizar o algoritmo prova de participação (PoS) para validar as transações, e tornar a rede mais confiável. O PoS tem como vantagem a baixa carga gerada na rede, logo, pode ocasionar menos atrasos na comunicação com a blockchain. Portanto, o autor apresenta um método que pode ser implementado para a construção de uma ferramenta que garanta o compartilhamento de imagens médicas de forma confiável e livre de adulterações. No entanto, o autor alerta os pesquisadores que desejam replicar o método, que se preocupem com a privacidade dos dados compartilhados, pois, essa proposta não garante esse requisito, que é importante para os sistemas de saúde.

4.2 Preservação da privacidade em ambientes blockchain para saúde

Discutindo especificamente os trabalhos que apresentam soluções para prover privacidade nos ambientes de saúde auxiliados pelo blockchain, é possível apresentar o trabalho de Guo *et al.* (2018), intitulado *Secure Attribute-Based Signature Scheme With Multiple Authorities for Blockchain in Electronic Health Records Systems*. Os autores propõem uma abordagem para a

troca de registros eletrônicos médicos, baseada em um esquema de multi autoridades (utilizando da Criptografia baseada em Atributos (CBA)), a fim de proteger a privacidade dos pacientes. Desse modo, os registros médicos são compartilhados e armazenados na blockchain, em que cada usuários necessita de um conjunto de assinaturas, utilizando a computação bilinear de *Diffie-Hellman*, para recuperar cada atributo. Utilizando essa técnica os autores afirmam que o sistema garante a privacidade, assim, há baixo risco de as informações dos pacientes serem violadas quando os dados são compartilhados, uma vez que, as chaves para acesso de cada atributo são distribuídas entre as autoridades. Após a definição da proposta, os autores realizaram a avaliação, que cobre a análise da segurança e do desempenho. Na análise da segurança uma abordagem teórica é seguida, com o propósito de provar matematicamente que os algoritmos de geração e validação das multi assinaturas são válidos e assim provam que o método é resistente a colisões. Na análise do desempenho, os autores provam que o custo para a geração e verificação das assinaturas são lineares, ou seja, possui baixo custo comparado com outras propostas na literatura. Além disso, o trabalho apresenta limitações a respeito dos dados criptografados que ainda estão sujeitos a perda ou vazamento de chaves, as informações criptografadas podem ter uma utilidade muito baixa para o usuário final e a identidade dos usuários nos dados compartilhados podem ser descobertas por ataques de re-identificação, como o de ligação.

Em [Al-Omar et al. \(2019\)](#) é discutido o ***Privacy-friendly platform for healthcare data in cloud based on blockchain environment***, ou ***MediBChain***, que emprega criptografia de Curvas Elípticas para garantir a privacidade dos pacientes. O foco da proposta é o gerenciamento de dados centrado no paciente, no qual somente ele consegue disponibilizá-los para outros provedores. Todos os registros médicos transitados pela rede são armazenados seguindo o esquema de nuvem, mas em um banco de dados baseado na blockchain. O projeto tem como principais contribuições: (i) garantia da segurança e privacidade fundamentada em responsabilidade, pseudônimo, autenticidade e integridade; (ii) blockchain permissionado para restringir usuários desconhecidos; (iii) os dados centrados no controle do paciente; e (iv) esquema de curvas elípticas para prover maior segurança e pseudônimos aos pacientes. Para o desenvolvimento da proposta os autores constroem modelos teóricos de contrato, em que se baseiam nas características da aplicação. Com essa base construída, os autores configuraram as simulações dos experimentos, seguindo a rede Ethereum e elaborando um contrato em Solidity, implementando, inclusive, o esquema de criptografia em Java (versão 1.8). Os resultados expressam que para o compartilhamento dos dados, o tempo de encriptação é diretamente proporcional ao tamanho da cifra. Mesmo a proposta oferecendo diversas vantagens, ainda tem limitações, que são: não é interoperável entre outras organizações de saúde; está suscetível ao roubo e perda de chaves, que podem causar a inutilidade de dados por conta do paciente ou médico; por fim, com a pseudo-anonimização os pacientes ainda estão sujeitos a ataques de ligação para inferir atributos reais com base em um conhecimento prévio sobre o paciente.

Ainda sobre a garantia de privacidade em sistemas de saúde baseados em blockchain, surge a proposta de [Gan et al. \(2020\)](#), nomeada ***Blockchain-based access control scheme with***

incentive mechanism for eHealth systems: patient as supervisor. O trabalho tem o propósito de proteger os direitos dos pacientes que utilizam os sistemas de saúde em nuvem, provendo privacidade e trazendo privilégios para que o próprio paciente gerencie seus dados a partir da blockchain. As contribuições que esse trabalho agrega aos sistemas de saúde são: (i) um método para o próprio paciente conceder acesso aos seus dados e assim os sistemas de saúde utilizarem legalmente os dados; (ii) controle de acesso para prover privacidade aos pacientes, pois somente os proprietários acessam seus dados; e (iii) incentivo para os pacientes compartilharem seus dados para pesquisa. Na elaboração da proposta, os autores desenvolveram um mecanismo de controle de acesso, que monitora os usuários pacientes, instituições médicas e terceiros. Assim, os pacientes podem compartilhar, acessar e revocar acesso às informações; as instituições de saúde podem acessá-las e gerar os blocos da rede; e os terceiros são capazes somente de acessar as informações. Os autores também adicionam um mecanismo de incentivo que recompensa os pacientes e instituições de saúde quando estes compartilham dados. Para o desenvolvimento da proposta, e execução dos experimentos foi empregada a rede Ethereum com a REMIX-IDE¹, em duas máquinas como mineradores e um contrato inteligente em Solidity. Os resultados para as simulações com 15 transações por segundo mostram sucesso, bem como melhorias na confiança e confidencialidade para compartilhar registros de saúde. O trabalho também apresenta algumas limitações, como o custo elevado para mineração dos blocos, em virtude de utilizar um protocolo PoW, assim como preocupações com de-anonimização dos dados compartilhados, que estão suscetíveis a ataques de ligação ou probabilísticos.

4.3 Sumário dos trabalhos

Para analisar os trabalhos relacionados, algumas características foram observadas, com base na proposta de Shi *et al.* (2020). O sumário dos trabalhos relacionados apresenta um resumo das contribuições e limitações da literatura frente à proposta desta pesquisa.

4.3.1 Visão geral

Para discutir as principais características de cada trabalho foi elaborada a Tabela 2 que evidencia algumas características que são importantes para preservar a segurança e privacidade em sistemas de saúde, bem como aqueles de que os sistemas baseados em blockchain necessitam. A Tabela 3, por sua vez, descreve uma discussão dos trabalhos com base na adaptação do modelo proposto por Shi *et al.* (2020), com o propósito de avaliar os requisitos que aplicações em saúde necessitam cobrir para serem mais seguras.

Com base no levantamento das características de cada trabalho coletado da literatura, foi possível realizar a discussão e a comparação das características desta proposta. Primeiramente, os ativos transferidos em cada trabalho são, em sua maioria, Registros Eletrônicos Médicos

---

¹ <https://remix.ethereum.org/>

Tabela 2 – Sumário das características levantadas dos trabalhos relacionados, que foram usadas para comparar como a nossa solução.

Trabalho	Características						
	Ativos armazenados	Método de armazenamento	Algoritmo de consenso	Plataforma de implementação	Modelo de privacidade	Rede blockchain	Nível de privacidade
(AZARIA <i>et al.</i> , 2016)	REM	Blockchain	PoW	Ethereum	Desconhecido	Não permissionado	Nenhum
(XIA <i>et al.</i> , 2017)	REM	Blockchain	PoW	Ethereum	Desconhecido	Não permissionado	Nenhum
(SHEN; GUO; YANG, 2019)	REM	Blockchain	BFT-SMaR	Implementação própria	Desconhecido	Permissionado	Nenhum
(PATEL, 2018)	Imagens	Blockchain	PoS	Implementação própria	Chaves pública e privada	Não permissionado	Nenhum
(GUO <i>et al.</i> , 2018)	REM	Blockchain	Desconhecido	Proposta teórica	CBA	Desconhecido	Protegido
(AL-OMAR <i>et al.</i> , 2019)	REM	Blockchain	PoW	Ethereum	CCE	Não permissionado	Protegido
(GAN <i>et al.</i> , 2020)	REM	Blockchain	PoW	Ethereum	Controle de acesso	Não permissionado	Protegido
Nossa solução	<i>Imagens</i>	<i>IPFS e blockchain</i>	<i>PBFT</i>	<i>Hyperledger fabric</i>	<i>K-Anonimato e Privacidade diferencial</i>	<i>Permissionado privado</i>	<i>Privado</i>

1) REM: Registros Eletrônicos de Médicos, 2) IPFS: Interplanetary File System, 3) PoS: Proof of Stake, 4) PoW: Proof of Work, 5) PBFT: Practical Byzantine Fault Tolerance, 6) CBA: Criptografia Baseada em Atributos, 7) CCE: Criptografia de Curvas Elípticas

(REM) – somente um deles, [Patel \(2018\)](#) trabalha na transferência de imagens. Portanto, para esta pesquisa foram utilizadas imagens médicas para serem compartilhadas em virtude de esse ser um ativo importante para a aquisição de conhecimento no campo da saúde ([GE *et al.*, 2012](#)).

No que se refere à característica método de armazenamento, que é a forma com que os dados são gravados, considerado as questões sobre segurança e privacidade, considerando os trabalhos de [Azaria *et al.* \(2016\)](#), [Xia *et al.* \(2017\)](#), [Al-Omar *et al.* \(2019\)](#) e [Gan *et al.* \(2020\)](#). Estes, armazenam seus dados utilizando somente da estrutura da blockchain do Ethereum, garantindo maior segurança e não repúdio aos dados, pois, são imutáveis. Os trabalhos de [Shen, Guo e Yang \(2019\)](#) e [Patel \(2018\)](#), por sua vez, também armazenam os dados em uma estrutura da blockchain, com uma implementação própria, contudo, não discutem esse aspecto com detalhes. Embora, as propostas apresentadas na literatura forneçam um método de armazenamento que garanta melhor segurança, elas ainda apresentam desafios relacionados com a privacidade e não se preocupam com a LGPD nesse quesito. Assim sendo, no que diz respeito ao armazenamento dos *logs* e metadados na blockchain (guardados no CouchDB) e os dados puros (imagens), neste trabalho estes são salvos na rede P2P *Off-chain*, configurada com a plataforma IPFS. Essa característica traz para a proposta a vantagem da possibilidade de apagar os dados, seguindo a LGPD com o direito ao apagamento dos dados ([BRASIL, 2018](#)).

Trabalhos da literatura utilizam da blockchain para armazenar e transferir seus dados, também fazem o uso de protocolos de consenso para estabelecer confiança entre os pares. Dessa forma, o protocolo utilizado em [Azaria *et al.* \(2016\)](#), [Xia *et al.* \(2017\)](#), [Al-Omar *et al.* \(2019\)](#) e [Gan *et al.* \(2020\)](#), é o PoW (ver na Seção 2.1.4) devido ao uso da plataforma Ethereum Mainnet, que aplica uma variação do PoW, em que há o processo de consenso não-permissionado. Em ([PATEL, 2018](#)) é aplicado o PoS não-permissionado, no qual há uma implementação própria. O trabalho de [Shen, Guo e Yang \(2019\)](#) utiliza um protocolo BFT ([CASTRO; LISKOV, 1999](#)), que tem o processo de consenso permissionado, sendo mais leve para estabelecer o acordo entre os

pares. [Guo et al. \(2018\)](#) elabora a própria rede, logo, não apresenta como funciona o protocolo de consenso. Para esta pesquisa foi utilizado o protocolo de consenso PBFT (veja na Seção 2.1.4), com o objetivo de consumir pouco recurso hardware e rede, além de ser mais seguro, tendo em vista que utiliza de um conjunto confiável para estabelecer o consenso entre os pares e validar os blocos.

Para a implementação de cada trabalho, foram utilizadas algumas plataformas que auxiliam na construção de contratos inteligentes, tal como da rede blockchain. Nos artigos de [Azaria et al. \(2016\)](#), [Xia et al. \(2017\)](#), [Al-Omar et al. \(2019\)](#) e [Gan et al. \(2020\)](#) a rede blockchain foi baseada no Ethereum, que possui facilidades para a implementação e melhor confiabilidade, no entanto, pode apresentar problemas com a privacidade dos pacientes, haja vista que, são redes públicas e que qualquer usuário pode ingressar para participar do consenso ([FENG et al., 2019](#)) – seguindo as regras de cada protocolo. Para os trabalhos de [Shen, Guo e Yang \(2019\)](#) e [Patel \(2018\)](#) foram elaboradas implementações próprias, respectivamente plataformas pública e privada. Em ([GUO et al., 2018](#)), não há implementação de plataforma, pois os autores propõem uma alternativa teórica e fazem testes teóricos para provar a efetividade do método. Portanto, neste trabalho, para a construção da rede e do contrato inteligente, foi utilizada a plataforma Hyperledger Fabric, que se baseia em uma rede permissionada privada, com contratos customizáveis na linguagem Go. Essa implementação fornece menor custo para a validação das transações entre os pares, como também facilidades para configurar o controle de acesso, bem como organizar o *framework* de forma privada.

As últimas características analisadas, são os modelos de privacidade utilizados e o nível de privacidade que cada proposta garante. Em relação aos níveis de privacidade foram divididos em: **nenhum**, quando não considerada privacidade ou considerada um método que não é efetivo contra violações; **protegido**, quando contempla um método de privacidade efetivo, mas ainda apresenta problemas com ataques de re-identificação; e **privado**, quando aborda um método efetivo contra violações da privacidade, que reduza riscos de ataques de re-identificação. Nos trabalhos de [Azaria et al. \(2016\)](#), [Xia et al. \(2017\)](#) e [Shen, Guo e Yang \(2019\)](#) os modelos de privacidade são desconhecidos e assim tem **nenhum** nível de privacidade, pois esse requisito não foi considerado pelos autores, somente a blockchain é utilizado para melhorar a segurança dos sistemas. O trabalho de [Patel \(2018\)](#) também apresenta **nenhum** nível de privacidade, pois o único método considerado para a proteção dos dados é o esquema de chave pública e privada. Além disso, os autores citam que para a replicação do sistema, devem ser consideradas as preocupações sobre privacidade. [Guo et al. \(2018\)](#), [Al-Omar et al. \(2019\)](#) e [Gan et al. \(2020\)](#) apresentam um nível de privacidade **protegido**, que pode ser mais adequado para os sistemas médicos, em virtude da manipulação de dados sensíveis. Na proposta de [Guo et al. \(2018\)](#) foi aplicada a técnica de CBA, em que cada atributo tem um controle de acesso para cada usuário da rede, porém, esta fornece um nível de privacidade protegido, uma vez que as chaves estão sujeitas a roubos ou perdas e a utilidade dos dados é baixa. [Al-Omar et al. \(2019\)](#) utilizam a CCE para prover a privacidade, a partir da pseudo-anonimização, no entanto, ainda está sujeito a ataques

de ligação de atributos para a re-identificação. Além disso, no trabalho de [Gan et al. \(2020\)](#), os pesquisadores propõem um esquema de controle de acesso, de modo a evitar acessos indevidos aos dados de saúde, no entanto este ainda apresenta problemas com privacidade envolvendo a re-identificação dos atributos.

4.3.2 Discussão

Com base nos levantamentos dos trabalhos relacionados, esta pesquisa tem o propósito de prover um nível maior de privacidade alicerçado em modelos que forneçam um nível privado. Esse nível, privado, descreve a utilização de métodos que sejam efetivos contra a violação da privacidade, mas também contra os ataques de re-identificação, utilizando os modelos K-Anonimato e Privacidade Diferencial. Os médicos são considerados usuários mais confiáveis, assim para eles acessarem foi empregado o modelo do K-Anonimato, que fornece defesas contras os ataques de ligação de registros, e tem maior utilidade dos dados para os dados finais. Os pesquisadores e profissionais que acessam os dados para utilizarem em pesquisas, recebem um resumo dos dados, pois estes são anonimizados seguindo o modelo de Privacidade Diferencial com maior entropia e menor utilidade, fornecendo um nível privado. Ambos os métodos aplicados têm o objetivo de mitigar ataques de re-identificação dos registros que estão na blockchain, para construir um sistema privado e seguro para os ambientes médicos. Ademais, o *framework* proposto por esta pesquisa disponibiliza um esquema de rastreamento dos dados compartilhados para mitigar os impactos do vazamento.

Tabela 3 – Requisitos dos sistemas para comparar as propostas que visam prover privacidade aos pacientes que utilizam a estrutura da blockchain em sistemas de saúde. Adaptado de [Shi et al. \(2020\)](#).

Trabalho	Requisitos						
	Privacidade	Anonimização	Integridade	Autenticação	Controle	Auditoria	Responsabilização
(AZARIA et al., 2016)			✓	✓	✓	✓	
(XIA et al., 2017)			✓	✓	✓	✓	
(SHEN; GUO; YANG, 2019)			✓	✓	✓	✓	
(PATEL, 2018)	✓		✓	✓	✓	✓	
(GUO et al., 2018)	✓			✓	✓	✓	
(AL-OMAR et al., 2019)	✓	✓	✓	✓	✓	✓	
(GAN et al., 2020)	✓		✓	✓			
Nossa Solução	✓	✓	✓	✓	✓	✓	✓

A Tabela 3 exibe o sumário dos principais requisitos para comparação dos trabalhos relacionados com esta pesquisa. Para tanto, foi adaptado o método de [Shi et al. \(2020\)](#), ao qual refere à análise dos requisitos que um sistema precisa cobrir para garantir maior segurança. Portanto, foram analisados os requisitos: privacidade, anonimização, integridade, controle, auditoria e responsabilização. Para cada uma das definições, seguimos alguns significados específicos. São eles: (i) privacidade, é o direito do proprietário de dados de controlar, bem como de manter sua confidencialidade ([BRASIL, 2018](#)); (ii) anonimização, é a utilização de modelos para eliminar o vínculo de uma pessoa natural com seus dados, de forma direta ou indireta ([BRASIL, 2018](#)); (iii) integridade, é a proteção contra a alteração identidade ou

eliminação indevida, que também garante o não repúdio (BARKER, 2003); (iv) controle, é a adição de técnicas e políticas para o gerenciamento de risco, com base no controle do fluxo dos ativos transferidos (GARFINKEL, 2015); (v) auditoria, é o processo de revisão dos ativos para analisar se as políticas estabelecidas estão sendo cumpridas (NIELES; DEMPSEY; PILLITTERI, 2017); e (vi) responsabilização, é definido no sentido da ação de rastreamento das atividades das organizações que tratam dados, também pode ser descrito como *Accountability* (NIELES; DEMPSEY; PILLITTERI, 2017).

Os trabalhos relacionados cobrem alguns dos requisitos levantados, no entanto, apresentam lacunas em relação a outros requisitos, como é o caso da anonimização e responsabilização. Para a anonimização, somente em Al-Omar *et al.* (2019) foi possível perceber preocupação com esse requisito, por outro lado, os estudiosos abordam somente a pseudo-anonimização a partir da blockchain. Referente ao requisito de privacidade, as propostas de Azaria *et al.* (2016), Xia *et al.* (2017) e Shen, Guo e Yang (2019) não consideram fortemente esse requisito como os outros trabalhos, assim, as características citadas pelos autores para auxiliar na preservação da privacidade, são baixas. Por fim, no que se refere ao trabalho de Gan *et al.* (2020) não são atendidos os requisitos de anonimização, controle, auditoria e responsabilização, devido ao fato da proposta ser uma modelagem teórica que se preocupa especificadamente com a segurança e privacidade. Em relação a anonimização é uma prática necessária para esse trabalho, pois mesmo o paciente aprovando o acesso, ainda podem ocorrer vazamentos por parte do paciente.

Em vista disso, o *framework* proposto por esta pesquisa atende aos requisitos da Tabela 3 com o objetivo de se ajustar com as regras da LGPD. Assim, reduz problemas com a privacidade dos pacientes, como também responsabiliza os agentes de tratamento pelo manuseio indevido dos dados compartilhados. Os pontos que se destacam dos outros trabalhos são a anonimização e responsabilização, bem como o nível privado. A anonimização se diferencia por aplicar modelos para prover melhor privacidade e reduzir riscos da re-identificação dos registros e mitigar ataques de ligação. Para a responsabilização utiliza do *token* de rastreio, em conjunto com a blockchain, para auditar os ativos, no caso de ocorrerem vazamentos. Os outros requisitos satisfeitos, são auxiliados pelo Hyperledger Fabric que disponibiliza uma estrutura customizável (que pode ser complexa de ser construído) para controle de acesso, autenticação, configuração dos contratos inteligentes e para auditar os registros.

FRAMEWORK PROPOSTO PARA A PRESERVAÇÃO DA PRIVACIDADE

O presente capítulo descreve o *framework* baseado em blockchain para preservar a privacidade no compartilhamento de dados de saúde contra os ataques de re-identificação. Foi empregado um método de privacidade, além de políticas de acesso para os usuários ingressarem na rede.

5.1 *Framework* proposto

Esta pesquisa, propõe um *framework* baseado em blockchain intitulado BCPSHield (BlockChain to Privacy-preserving in SHaring sensItive hEaLth Data) que estrutura uma rede de compartilhamento de dados de saúde para prover a privacidade dos pacientes e protegê-los de ataques de re-identificação, além de auxiliar na responsabilização por possíveis vazamentos, garantindo o não repúdio. A Figura 4 descreve a visão geral do *framework*. Ademais, o código completo da proposta pode ser encontrado no Github¹.

Cada componente do *framework* atua para fornecer privacidade e rastrear os dados compartilhados, assim, as próximas seções apresentam todos eles de forma detalhada. Para formar o *framework*, temos os componentes que gerenciam os processos, que são: **componente do usuário, componente do provedor de saúde, componente de privacidade e rede blockchain**. Cada provedor de saúde interage com os componentes do provedor de saúde e de privacidade, que são executados localmente, a partir de dois *Docker containers*. Dessa forma, eles se comunicam na rede local e transferem dados para outras partes do *framework* utilizando uma *Application programming interface* (API), em português é a Interface de Programação de Aplicação.

---

¹ <https://github.com/eriksonJAguilar/bcshield-framework>

Figura 4 – Visão geral do *framework* BCPSHield, que é composto por 11 fluxos para compartilhar e recuperar dados. Os fluxos são: (1) aplicação arbitrária do usuário; (2) interface de comunicação com os demais componentes; (3) canal de envio de dados; (4) observador local; (5) base de dados local; (6) componente de privacidade dos dados; (7) inserir *token* de rastreamento nos dados compartilhados; (8) rede *Off-chain*; (9) rede blockchain; (10) canal de envio dos dados compartilhados; (11) canal de troca de informações compartilhadas com a interface.

5.1.1 Descrição dos componentes do framework

Esta seção descreve o funcionamento e implementação dos componentes do *framework* proposto, com o objetivo de apresentar melhor suas funcionalidades. Em vista disso, o propósito deste trabalho é **desenvolver um framework baseado em blockchain para preservar a privacidade no compartilhamento de dados de saúde e reduzir os riscos a ataques re-identificação**. Em complemento foi explorada a seguinte questão de pesquisa: “*Como mitigar os riscos à privacidade dos pacientes no compartilhamento de dados de saúde em redes baseadas em blockchain?*”.

Componente do usuário

O componente do usuário tem o propósito de fornecer uma área de interação entre os usuários que seja de fácil integração em diferentes aplicações para a utilização do *framework*. Desse modo, para a sua construção foi utilizada a linguagem *JavaScript* com o *Express*, a partir de métodos baseados no padrão *Representational state transfer* (REST). Para que os usuários possam desenvolver aplicações, consumir esta API e assim, utilizarem de todas as funcionalidades do BCPSHield em outras aplicações. Ademais, a API também facilita o acesso à rede blockchain, uma vez que utiliza o SDK do Hyperledger Fabric, que proporciona métodos para acessarem os atributos do contrato inteligente. Portanto, os métodos utilizados na blockchain são para armazenar *logs* e metadados, consultar dados e auditar vazamentos.

Componente do provedor de saúde

O componente do provedor de saúde tem o objetivo de estabelecer a comunicação dos usuários finais com a rede blockchain e os provedores de saúde que armazenam os dados dos pacientes. O componente é constituído de três fragmentos: (i) a base de dados locais, que armazenam os dados dos pacientes; (ii) o controlador, que verifica quando há novas requisições; e o (iii) conector para o componente de privacidade. O fragmento principal deste componente é representado pelo controlador, que gerencia o tráfego das requisições e transações. Além disso, este componente verifica novas requisições, captura os dados solicitados das respectivas bases locais e os envia para serem anonimizados através do conector, que comunica com componente de privacidade. Após anonimizados, os dados são enviados para o controlador novamente, que os encaminha pelo canal de comunicação para a rede blockchain.

Para a implementação desse componente foi empregada a linguagem Go, pois esta apresenta bom suporte e facilidades para desenvolver aplicações distribuídas. Dessa forma, para integrar as partes, foram aplicados *Docker containers* que auxiliam na instalação da aplicação e podem ser executados em qualquer máquina, desde que o *Docker* esteja instalado. Em suma, cada provedor de saúde tem um módulo individual sendo executado localmente, eliminando, assim, algumas preocupações relacionadas com a segurança, como a confidencialidade.

Componente de privacidade

Para prover privacidade aos usuários quando compartilham seus dados, foi proposto este componente, uma vez que ele anonimiza os dados localmente antes de disponibilizá-los para pesquisadores ou profissionais da saúde. Assim, é apresentado um método de privacidade, que considera o papel do usuário na rede para escolher o método de privacidade adequado. Se o usuário for um pesquisador, ele irá acessar um resumo dos dados para auxiliar em seus estudos, isto é, com a Privacidade Diferencial. Já o médico que trata um paciente consegue acessar os dados aplicando menos restrições, mas anonimizado com o K-Anonimato, que tem os atributos anonimizados mais expressivos. Posto isso, nas próximas seções serão discutidos com mais detalhes os níveis de acesso e os algoritmos aplicados.

Os métodos utilizados para proteger os dados foram o K-Anonimato e a Privacidade Diferencial, de modo a prover a privacidade das informações compartilhadas e garantir que estas sejam visíveis a somente um subconjunto de entidades (AXON; GOLDSMITH; CREESE, 2018). Esses modelos trabalham com dados numéricos e categóricos (que representam símbolos como: CEP, sexo e tipo sanguíneo), em que os algoritmos aplicam diferentes métodos para anonimizar os dados, principalmente quando se trata dos categóricos. Esse módulo se comunica com o controlador e decide quais são os modelos de privacidade aplicados, depois disso o controlador passa um sinal que indica se o dado é numérico ou categórico, tal como o nível de permissão do usuário que irá acessá-lo.

Rede blockchain

O componente da blockchain armazena os *logs* e metadados, fazendo intersecção com a

rede *Off-chain*, que armazena o dado puro fora da blockchain. Quando o dado é recebido no canal de entrada dessa camada, ele é dividido em duas partes, as imagens DICOM e o metadado, que são os atributos da imagem. Antes de armazenar o dado no *Off-chain* é gerado um *token* único com a seguinte característica: $token = sha256('PATIENT_{ID}' + 'TIMESTAMP_{ACESSO}')$, que é um valor *hash* com 32 bytes. Os valores inseridos para gerar o *hash* definem uma configuração inicial, que pode apresentar vulnerabilidades devido ao *Timestamp* ter baixa entropia e o identificador do paciente ser fixo. Para mitigar esse problema, o *token* pode ser construído da seguinte forma: $token = sha256('CONTEUDO_{imagem}' + 'TIMESTAMP_{acesso}' + 'Numero')$; em que é utilizado o conteúdo interno da imagem, o *timestamp* e um número pseudo-aleatório. Esse *token* é inserido nos metadados da imagem, com o arquivo que é armazenado no IFPS, tornando-a única, com o propósito de auxiliar na identificação de vazamentos. Além disso, O fluxo de inserção do *token* (ver a Figura 17) e de auditoria de imagens vazadas (ver a Figura 18) podem ser observados no Apêndice A.

Após inserir o *token*, os dados são armazenados no *Off-chain*, que é construído com o *framework InterPlanetary File Systems* (IPFS) e tem uma API para conexão, que foi construída, por sua vez, na linguagem Go. Os provedores de saúde enviam os dados para esta rede, recebendo como resultado uma chave representada por um valor *hash*, que é utilizado para recuperar os dados. Com a integração de uma rede fora da blockchain, é possível remover os dados do *framework*, eliminando apenas a referência da imagem, contribuindo, assim, para garantir o direito a remoção, segundo a LGPD. Além disso, com o *Off-chain* é possível armazenar dados que ocupam grande espaço.

Para construção da blockchain foi utilizado o Hyperledger Fabric, que é composto por 3 organizações com 2 pares cada, um ordenador e um banco de dados *CouchDB* replicado e distribuído para cada nó. A rede do HLF tem o objetivo de armazenar os *logs* e metadados das imagens, pois se trata de uma rede livre de adulteração. Também foi definido um contrato inteligente que contém as regras de negócio para armazenar e recuperar os dados, seguindo as políticas de acesso e de privacidade. Por fim, a Figura 5 representa uma visão do *framework* HLF para esta proposta, em que contém 1 *ordering service node* (OSN), 3 autoridades certificadoras (CAs) uma para cada organização e 6 pares, no qual cada organização contém 2. Para comunicar com o HLF é possível acessar a API que utiliza do SDK do *Fabric* para comunicar com os pares. O fluxo de comunicação da rede blockchain do HLF inicia pela API que comunica com o OSN, estabelecendo o contato com o par através do canal *gossip*. Contudo, antes de conectar o cliente com o par de sua organização, há uma verificação por parte das autoridades certificadoras, para, assim, definir a comunicação com os pares e os bancos de dados que armazenam seus dados.

Em suma, a Figura 5 pode representar o compartilhamento de imagens entre pesquisadores e hospitais (que são os provedores de saúde) que realizam exames de imagens. O paciente precisa aceitar o compartilhamento, pois os estudos contêm seus dados. Para o compartilhamento são utilizados os canais de comunicação, que passam pelo protocolo de *gossip* para acessarem os

Figura 5 – Arquitetura do Hyperledger Fabric para a rede blockchain desta proposta. Baseado em [Androulaki et al. \(2018\)](#).

nós que tem os dados armazenados. As organizações configuradas foram divididas em: paciente (PA), pesquisador (PE) e provedor de saúde (HP); cada um das organizações tem um papel na rede do HLF. Por fim, as APIs são utilizadas para a interação dos clientes com a rede, bem com os métodos que foram implementados no contrato inteligente.

5.1.2 Descrição do fluxo de dados do framework

Os fluxos que representam os dados trafegados no *framework* são ilustrados na Figura 4 pelos números de 1 a 11 em verde. A seguir são apresentados como cada fluxo faz a interação entre os componentes da rede:

1. **Aplicação do usuário:** este fragmento pertence ao componente do usuário e tem o objetivo de fornecer um meio de comunicação amigável entre os usuários (profissional da saúde, paciente e pesquisador);
2. **Interface de comunicação:** é uma API REST que faz a comunicação com os outros componentes da rede. Para seu desenvolvimento, foi utilizada a linguagem *JavaScript* com o *framework Express*, além do *Software development kit (SDK)* do Hyperledger Fabric;
3. **Canal de envio de dados:** é o canal em que os dados transitam e enviam as requisições dos usuários aos demais componentes da rede. Esse fragmento fica observando a chegada de novas requisições pelo canal de comunicação e notifica os demais componentes;
4. **Controlador:** pertence ao componente do provedor de saúde, é executado localmente por um *Docker container*. Tem o objetivo de observar as requisições encaminhadas pelo canal de comunicação, como também coletar os dados solicitados da base local. Por fim, realiza o envio e recebimento de dados que o componente de privacidade necessita tratar;
5. **Base de dados local:** representa a base de dados, na qual o provedor de saúde armazena os dados de seus pacientes;

6. **Privacidade dos dados:** é o componente que aplica privacidade aos dados antes destes serem compartilhados, utilizando dois métodos, a Privacidade Diferencial e o K-Anonimato. Também vale ressaltar que, esse processo é realizado localmente em segundo plano com o módulo do provedor de saúde e é aplicado aos dados;
7. **Token de rastreamento:** os dados e metadados enviados pelo componente do provedor de saúde são recebidos e divididos em dados e metadados. Os metadados são extraídos da imagem DICOM e é adicionado um *token*, que pode contribuir no rastreamento de dados compartilhados. Os dados são as imagens que são armazenados no *Off-chain*;
8. **Rede Off-chain:** é a rede que compõe os repositórios para os dados puros, como as imagens, por exemplo. A rede *Off-chain* também auxilia na futura remoção dos dados, para, assim, seguir os requisitos de controle de acesso da LGPD;
9. **Anonimização dos dados:** os registros anonimizados e *logs* de acesso são inseridos na rede blockchain. Desse modo, esses componentes são imutáveis e auxiliam na prestação de contas por parte das organizações pelos dados acessados;
10. **Canal de envio de dados compartilhados:** é o canal que os dados compartilhados transitam após o processamento;
11. **Canal de troca de informações compartilhadas com a interface:** aguarda os dados cujo processo foi solicitado, para, assim, enviá-los para a aplicação aos qual os usuários interagem.

Vale destacar que, os fluxos são importantes para definir como os dados compartilhados transitam na rede e quais componentes eles acessam. Para encaminhar os dados entre os componentes e verificar quais as ações são necessárias, o controlador é acionado, uma vez que, o mesmo envia para anonimizar, coleta dados do banco, envia para aplicar a privacidade e encaminha para o armazenamento na blockchain. Essa tarefa é importante para restringir o acesso a determinados ativos e evitar violações.

5.2 Framework proposto frente a arquitetura de referência

O *framework* proposto também foi descrito em baixo nível, seguindo a arquitetura referência definida por Yuan e Wang (2018). Para o *framework* proposto as camadas da arquitetura original foram restruturadas, adicionando a camada de privacidade e a camada de *Off-Chain*. A Figura 6 ilustra cada camada da arquitetura, que é análoga aos módulos definidos para o *framework* na seção anterior. O modelo de Yuan e Wang (2018) é uma ilustração teórica de como

Figura 6 – Arquitetura de referência das camadas da blockchain. Adaptado de [Yuan e Wang \(2018\)](#). As siglas apresentadas nas figuras são: K-Anonimato (KA), Criptografia Baseada em Atributos (CBA), Privacidade Diferencial (PD), *Mixing* (MX), *JavaScript Object Notation* (JSON), *Hyperledger Fabric* (HLF), *Certificate Authority* (CA), *Distributed Hash Table* (DHT), e *InterPlanetary File System* (IPFS).

a blockchain pode ser organizada em camadas. Desse modo, para esta proposta foi empregado o mesmo modelo, mas adicionado a *Off-chain* e a privacidade.

As camadas que caracterizam a arquitetura são ([YUAN; WANG, 2018](#)): (i) **camada de aplicação**: interage diretamente com os usuários a partir da API, que se comunica com uma aplicação, no caso deste trabalho é para área da saúde; (ii) **camada de privacidade**: descreve o processo de anonimizar os dados e a identidade dos usuários que se comunicam com a camada de contrato; (iii) **camada de contrato**: representa as regras de negócio definidas nos contratos inteligentes para moldar a lógica das transações; (iv) **camada de consenso**: define o protocolo de consenso utilizado para validar as transações da rede; (v) **camada de rede**: representa a rede blockchain, e define como os dados serão encaminhados e verificados na rede P2P. Em conjunto com a camada de rede e fora da blockchain, está a **a Off-Chain**, que é um modelo de rede sobreposta alicerçada em uma rede P2P, para armazenar dados sensíveis e seguir as políticas da LGPD; (vi) **camada de dados**: é composta pelos atributos do bloco que serão incluídos no livro-razão. Nessa camada os dados são criptografados para passar pelas outras camadas, assim, é aplicada a criptografia de curvas elípticas utilizando a curva *prime256v1* e SHA256 ([FABRIC, 2020c](#)).

Com base na arquitetura de referência a Figura 6(b) ilustra as camadas do *framework* em baixo nível de forma específica para esta proposta. Primeiramente é apresentada a **Camada de Aplicação** para os usuários acessarem uma aplicação que está conectada a uma API, para estabelecer a comunicação com a blockchain, dessa forma, eles podem fazer o compartilhamento

dos dados de saúde. A **Camada de privacidade** é um dos diferenciais para este em relação ao modelo de referência de [Yuan e Wang \(2018\)](#), pois reúne dois modelos de anonimização com a blockchain, para, assim, reduzir riscos a privacidade dos usuários. Os modelos selecionados foram: o K-Anonimato e a Privacidade Diferencial. A **Camada de Contrato**, por sua vez, apresenta o *chaincode* (contrato para o HLF) na linguagem Go, além das regras de negócio e algoritmos para definir a lógica no compartilhamento dos ativos. A **Camada de Contrato**, por sua vez, apresenta o *chaincode* (contrato para o HLF) na linguagem Go, além das regras de negócio e algoritmos para definir a lógica no compartilhamento dos ativos. A **Camada de Rede** compreende os componentes da blockchain do HLF, que tem os pares para comunicação, ordenadores para gerenciar a comunicação e as autoridades certificadoras (CA) para autenticar os usuários na rede. Cada componente foi descrito com mais detalhe na fundamentação teórica (ver Capítulo 2). Outra camada que se difere do modelo convencional é a **Camada de Off-chain**, possibilitando a adição de dados fora da blockchain, além proporcionar a remoção dos dados com o propósito de seguir os requisitos da LGPD. Na **Camada de Consenso** é integrado o protocolo de consenso PBFT, que é fundamentado no problema dos generais Bizantinos (ver Capítulo 2) e em nossa implementação do HLF este foi selecionado. A última é a **Camada de dados** composta pelo bloco de dados, que se complementando com o JSON – visto que é o tipo de dado tratado pelo HLF. Também nessa camada é utilizado o *Elliptic Curve Digital Signature Algorithm* (EDSA) com a curva *prime256v1*, para assinar as transações enviadas e proteger os ativos envolvidos ([FABRIC, 2020c](#)).

5.3 Controle de acesso

O objetivo do estabelecimento das políticas de privacidade e da construção da rede blockchain privada é fornecer níveis de acesso para proteger os dados dos pacientes. Com as políticas de controle de acesso é possível mitigar o acesso indevido aos dados e controlar quem acessa determinado ativo. Portanto, a Tabela 4 descreve os métodos de escrita e leitura aos componentes do *framework* que apresentam uma visão geral das políticas de acesso, focando nos modelos de privacidade, rede blockchain, *Off-chain* e a base de dados local. Além disso, o controle de acesso é definido para os usuários paciente, pesquisador e profissionais da saúde (médicos, gestores, enfermeiros), bem como para o provedor de saúde que disponibiliza os dados.

Os usuários da rede são: pesquisadores, profissionais da saúde e pacientes. O provedor de saúde também pode ser considerado um usuário, pois são organizações confiáveis que trocam dados no *framework*. Os profissionais de saúde e pesquisadores podem acessar dados utilizando os métodos K-Anonimato ou Privacidade Diferencial para leitura, obtendo como resultado respectivamente dados com parte dos atributos ocultos, ou um resumo dos dados a partir de um histograma com resultados estatísticos. A blockchain é acessada pelos provedores de saúde que são portadores dos dados dos pacientes, podendo executar ações de leitura e escrita. O paciente também pode fazer a leitura e atualização dos dados na blockchain, com o objetivo de aprovar

Tabela 4 – Políticas de controle de acesso para os métodos de privacidade e componentes do *framework*. Também apresenta os usuários que compõem a arquitetura e as ações que os mesmos podem realizar.

	K-Anonimato	Privacidade Diferencial	Blockchain	Off-Chain	Base local de dados
Provedor de saúde			Leitura e escrita	Leitura e escrita	Leitura
Pesquisador	Leitura	Leitura	Leitura	Leitura	
Profissional de saúde	Leitura	Leitura	Leitura	Leitura	
Paciente			Leitura e atualização	Leitura	

o compartilhamento dos dados atualizando o atributo que o usuário aceitou para verdadeiro. A rede *Off-chain*, é acessada por todos os usuários do *framework*, em que o provedor de saúde adiciona os dados e os demais usuários podem recuperá-los. Por fim, o provedor de saúde pode acessar a base de dados local, que representa um exame pertence a um paciente, armazenado em um hospital, e que poderá compartilhar com um pesquisador ou profissional de saúde.

5.4 Modelo de privacidade

Como discutido anteriormente, tendo como base os níveis de acesso e os papéis dos usuários, dois mecanismos de privacidade foram definidos: o K-Anonimato e a Privacidade Diferencial. Para anonimizar os dados com a Privacidade Diferencial foi elaborado um algoritmo seguindo os fundamentos apresentados no estado da arte, como em [Dwork e Roth \(2014\)](#). Para a implementação do algoritmo foi utilizada a linguagem Go para o desenvolvimento de um módulo que pode ser estendido para outros projetos e foi publicado no Github². Vale ressaltar que, as consultas aplicadas nesse modelo são para a média e o histograma, que respectivamente são utilizadas para valores numéricos e categóricos. O Algoritmo 1 descreve o fluxo lógico para o funcionamento da Privacidade Diferencial empregada nesta pesquisa. Um ponto a se destacar é que, os modelos de privacidade foram construídos a partir de uma implementação própria, sem utilizar bibliotecas. Na literatura existem bibliotecas eficientes para ambos os métodos, porém, para o cenário desta pesquisa o funcionamento se tornou complexo, visto que, quando empregada em conjunto com a blockchain o contrato emitiu diversos erros, logo, foi necessária uma implementação própria.

O algoritmo 1 recebe como entrada um conjunto de dados D e um nível de ruído ϵ (linha 1) e tem como saída o dado anonimizado D' (linha 2) que corresponde a uma consulta com privacidade. Primeiro, é declarada uma variável Q (linha 3) com valor nulo, que é proporcional a consulta que será anonimizada, isto é, a consulta corresponde ao D agregado, pela média ou pelo

² <https://github.com/eriksonJAguilar/godiffpriv>

histograma. Com a variável Q declarada o próximo passo foi verificar o tipo do dado D (linha 4), sendo numérico ou categórico, uma vez que, esse processo é importante para definir a consulta que será empregada para agregar os valores. A partir da seleção do tipo do dado o algoritmo recupera um conjunto A (linha 5) que corresponde a amostra que o dado D pertence, ou seja, uma coluna do banco de dados, para, assim, realizar a consulta e agregar os dados (linha 6 ou linha 9, dependendo do tipo de dado). Os valores de A são úteis para complementar o algoritmo da Privacidade Diferencial, visto que se calcula a sensibilidade Δf (linha 11), que se refere ao impacto da remoção de uma única linha da base em relação a uma base de dados B original e uma B' similar. Posteriormente a ponderação do Δf é calculado o índice b (linha 12) que é a razão entre a sensibilidade e o nível de ruído ϵ , implicando no desvio da distribuição. Esse índice é empregado para gerar o ruído aleatório (linha 13) fundamentado na distribuição de probabilidade Laplaciana, resultando na variável *ruído* que é somada a consulta Q (linha 14). Desse modo, D' é obtido que representa a consulta com ruído para o algoritmo (linha 15).

Algoritmo 1 – Algoritmo Privacidade Diferencial modificado

```

1: Entrada: (dado a ser anonimizado  $D$ ,  $\epsilon$ )
2: Saída: dado anonimizado  $D'$ 
3:  $Q \leftarrow \text{Nulo}$  ▷ Define a consulta
4: se  $\text{tipo}(D) = \text{numerico}$  então
5:    $A \leftarrow \text{recupera amostras da base de dados}$ 
6:    $Q \leftarrow \text{Media}(A)$ 
7: senão
8:    $A \leftarrow \text{recupera amostras da base de dados}$ 
9:    $Q \leftarrow \text{Histograma}(A)$ 
10: fim se
11:  $\Delta f \leftarrow \text{Sensibilidade}(A)$ 
12:  $b \leftarrow \Delta f / \epsilon$  ▷ calcula desvio da distribuição
13:  $\text{ruído} \leftarrow \frac{1}{2b} \text{Exp}(-\frac{|\epsilon|}{b})$  ▷ gera o valor aleatório
14:  $D' \leftarrow Q + \text{ruído}$ 
15: retorna  $D'$ 

```

Para o K-Anonimato foi desenvolvido um algoritmo para anonimizar dados numéricos e categóricos, seguindo as abordagens de supressão e generalização com $K = 4$. O valor do K foi escolhido de forma arbitrária a partir da literatura como em [Machanavajjhala et al. \(2007\)](#), vale destacar que, não há um método efetivo para a escolha o K ([DEWRI et al., 2008](#)). Na supressão para dados categóricos, são inseridos asteriscos (*) no final da sentença, a fim de reduzir a compreensão do significado da palavra. Para valores numéricos, é utilizada a generalização com base no agrupamento dos valores, por exemplo, o atributo *idade* = 17 anos poderia ser generalizados para *idade* = de 10 até 20 anos. Na construção da generalização numérica, é aplicada a regra de *Sturges* ([SCOTT, 2009](#)) que auxilia na divisão dos grupos de um conjunto de dados, é útil na construção de histogramas. Nesse projeto foi empregado no procedimento de generalização para dados numéricos. Em suma, o Algoritmo 2 define como esse modelo funciona.

O algoritmo 2 recebe como entrada um conjunto de dados D (linha 1) e tem como retorno o dado anonimizado D' (linha 2). Para aplicar a anonimização é considerado o $K = 4$, tais como valores categóricos e numéricos, para, assim, aplicar respectivamente os algoritmos de supressão e generalização. Para dados categóricos, é aplicado o algoritmo da supressão que substitui os últimos valores da *string*, com auxílio de uma expressão regular. Em dados numéricos, é necessário generalizar os valores do conjunto de dados, que nesse caso foram recuperados e armazenados na variável A (linha 6), que representa a coluna corresponde aos valores na base. Desse modo, foi empregado o método de *Sturges* (linha 8), para dividir a coleção de valores numéricos em classes, sendo k o número de classes, n o tamanho do conjunto de dados, que são variáveis essenciais para o cálculo. No processo de cálculo é criada uma lista C e nela são adicionados os limites inferior e superior primários (linha 13), que são modificados para cada classe, até que seja alcançado o último valor do conjunto de dados (da linha 15 a 20). Cada valor dentro do conjunto será modificado para um intervalo específico seguindo as classes e os limites inferior e superior. Por exemplo, considerando valores arbitrários $X = \{10, 5, 45, 19, 60\}$, sendo $x \in X$, além dos limites para a classe 1, sendo o inferior $l_i = 0$ e o superior $l_s = 20$. Os valores de X que correspondem a classe 1 são: $x_1 = 10$, $x_2 = 5$, $x_4 = 19$, os demais valores pertencem a outras classes.

Algoritmo 2 – Algoritmo K-Anonimato modificado

```

1: Entrada: dado a ser anonimizado  $D$ 
2: Saída: dado anonimizado  $D'$ 
3: se  $\text{tipo}(D) = \text{categorico}$  então
4:    $D' \leftarrow D.\text{expRegular}(".4)/s", "*****")$ 
5: senão
6:    $A \leftarrow \text{recupera amostras da base de dados}$ 
7:    $n = \text{tamanho}(A)$  ▷ número de amostras
8:    $k = 1 + \log_2(n)$  ▷ quantidade de grupos para divisão
9:    $\text{limite}_{sup} \leftarrow \max(A)$ 
10:   $\text{limite}_{inf} \leftarrow \min(A)$ 
11:   $\text{limiar} \leftarrow (\text{limite}_{sup} - \text{limite}_{inf})/k$ 
12:   $C \leftarrow \text{Nulo}$ 
13:   $C \leftarrow \text{adiciona}((\text{limite}_{sup}, \text{limite}_{inf}))$ 
14:   $i \leftarrow 1$ 
15:  enquanto  $i < k$  faça
16:     $\text{limite}_{inf} \leftarrow \text{limite}_{sup} + 1$ 
17:     $\text{limite}_{sup} \leftarrow (\text{limite}_{inf} + \text{limiar}) + 1$ 
18:     $C \leftarrow \text{adiciona}((\text{limite}_{sup}, \text{limite}_{inf}))$ 
19:     $i \leftarrow i + 1$ 
20:  fim enquanto
21:  enquanto  $j < \text{tamanho}(D)$  faça
22:    se  $D[j] \leq \text{limite}_{inf}$  e  $D[j] \geq \text{limite}_{sup}$  então
23:       $D'[j] \leftarrow \text{limite}_{inf} + \text{limite}_{sup}$ 
24:       $j \leftarrow j + 1$ 
25:    fim se
26:  fim enquanto
27: fim se
28: retorna  $D'$ 

```

EXPERIMENTOS E RESULTADOS

Este capítulo apresenta a aplicação do *framework* em um cenário de simulação e define como os experimentos foram conduzidos, bem como seus respectivos resultados. As características avaliadas foram o desempenho do hardware e rede para a blockchain e para o módulo do provedor de saúde. Além disso, é descrito o cenário de aplicação, a avaliação dos requisitos de privacidade e as discussões sobre os resultados. Portanto, este capítulo demonstra que a blockchain tem um custo viável para ser aplicado aos sistemas saúde. O capítulo também discute a efetividade do *framework* no rastreamento dos ativos compartilhados a partir do *token*, além de resistir à colisões, que visam invalidar o *hash*. Em suma, são realizadas as simulações para os requisitos de privacidade com base no cenário de aplicação, que verificam se o *framework* segue os requisitos listados pelas metodologias *Open Web Application Security Project* (OWASP) e *Privacy by design* (PbD). Assim, são realizadas as análises dos experimentos para os níveis de privacidade (quantidade de entropia) e utilidade dos dados (distância entre o dado original e o anonimizado), com o intuito de mostrar a estabilidade dos algoritmos propostos na garantia da privacidade do paciente.

6.1 Configuração dos experimentos

Para realizar as simulações, utilizamos máquinas virtuais na *Google Cloud Platform* (GCP). Nos experimentos com a blockchain, as máquinas utilizadas têm a seguinte configuração: 9 GB de RAM, 2 CPUs virtuais, 1 TB de disco e o sistema operacional Linux Ubuntu 18.10. Essa máquina suporta diversos *Docker containers* que auxiliaram na execução do Hyperledger Fabric. O processo de avaliação da blockchain, foi seguido fundamentado alguns trabalhos da literatura, tal como o de [Tanwar, Parekh e Evans \(2020\)](#), que emprega o Hyperledger Caliper para facilitar a configuração e execução dos experimentos. Na avaliação de desempenho dos componentes do provedor de saúde, que executam e se comunicam a partir de *Docker containers*, esses componentes foram executados em um cliente com 8 GB de RAM, processador Core i5

2.4 GHz, armazenamento *flash* de 256 GB e sistema operacional MacOS Catalina 10.15. Além disso, para acessar os dados na rede *Off-chain*, tais componentes foram configurados em uma máquina virtual na GCP, com 3 GB de RAM, 2 CPUs virtuais, 1 TB de disco e Linux Ubuntu 18.10, com o IPFS instalado, essa máquina foi o nó central da rede P2P, além de integrar outros nós distribuídos para os testes.

Os dados manipulados nos experimentos foram sintéticos, pois não foram encontradas bases de imagens que contenham os dados sensíveis dos pacientes para serem anonimizados. Para contornar a carência por dados, foram gerados dois conjuntos sintéticos a partir de registros médicos públicos (com imagens e registros de texto). Em vista disso, os dois conjuntos sintéticos foram divididos em metadados e imagens no formato *Digital Imaging and Communications in Medicine* (DICOM). Esse processo foi realizado em virtude de as imagens públicas adquiridas apresentarem alguns atributos anonimizados, dessa forma, não era possível utilizá-las nos experimentos, sendo necessária a substituição por metadados sintéticos. Para gerar os metadados, seguimos como modelo o conjunto de registros médicos do estado de Nova Iorque, que engloba dados pessoais de vários provedores de saúde (DOH, 2016), totalizando 3,84 GB, no qual foram utilizados 5.000 registros nos experimentos. As imagens aplicadas nos experimentos foram recuperadas da base *The Cancer Image Archive* (TCIA), com a coleção *Clinical Proteomic Tumor Analysis Consortium* composta por 52.072 arquivos de imagens, totalizando 28,8 GB, em que foram utilizadas 5000 imagens nos experimentos (CLARK *et al.*, 2013).

6.2 Descrição do estudo piloto e cenário de ataque

A motivação para a utilização de imagens médicas nesta pesquisa, é que estas são importantes para o diagnóstico de diversas doenças e são utilizadas em diversos cenários para potencializar a área da saúde (GE *et al.*, 2012). Com a distribuição dessas imagens é possível ampliar o campo de estudo e contribuir para pesquisas em saúde e em outras áreas (GE *et al.*, 2012). No entanto, sistemas de saúde que armazenam imagens apresentam problemas relacionados com a segurança, exibindo uma arquitetura centralizada com um ponto único de falha (LANGER *et al.*, 2015). Desse modo, esses sistemas ficam vulneráveis a ataques com o propósito de causar a indisponibilidade, como exemplo o *Denial-of-service* (DoS), ou até mesmo vaziar dados. Visando à proteção desses sistemas, pode-se aplicar a blockchain para mitigar riscos no compartilhamento de imagens em sistemas de saúde para pesquisa. Esta tecnologia garante aos sistemas a descentralização do armazenamento das imagens, isto é, com base em um livro-razão replicado e distribuído, sem ponto único de falha (GREVE *et al.*, 2018).

Mesmo a tecnologia blockchain provendo benefícios a segurança dos sistemas computacionais ainda apresenta desafios envolvendo privacidade (Jin *et al.*, 2019). Assim, esses sistemas estão suscetíveis a ataques re-identificação, em que os registros dos usuários podem ser ligados à sua identidade real com auxílio de uma base de dados complementar (KUO; OHNO-MACHADO,

2018). A partir desses problemas, surgem as regulações de privacidade para legitimar a forma de tratamento dos dados, como HIPAA, que é específica para dados de saúde, além da LGPD, a lei brasileira (Jin *et al.*, 2019).

A partir das problemáticas apresentadas, para esta pesquisa foi elaborado um estudo piloto para o compartilhamento de imagens médicas entre pesquisadores, médicos e pacientes. Neste projeto, os pacientes controlam os seus dados, podendo compartilhar os exames com seus médicos, mas também com organizações de pesquisa com o intuito de contribuir com a ciência. Cada médico aguarda o recebimento das imagens para sua análise, seja para realizar algum diagnóstico aprofundado ou um exame de rotina. Os pesquisadores solicitam estudos para as organizações de saúde, que pedem autorização aos pacientes para, assim, compartilharem as imagens. O cenário para a pesquisa se resume em um pesquisador que solicita imagens para estudos sobre reconhecimento de padrões, por exemplo, desse modo o paciente recebe a solicitação e decide se compartilha ou não o seu exame para pesquisa, a Figura 7 descreve esse cenário. As preocupações dessa aplicação estão relacionadas à privacidade dos dados do paciente ou até mesmo aos vazamentos, nesse sentido, é necessária a integração de outras tecnologias para mitigarem esses problemas. Portanto, o estudo piloto aplica o *framework* **BCPShield** para reduzir os riscos a violações de privacidade no compartilhamento de imagens médicas.

Figura 7 – Cenário de estudo utilizado como base para as simulações. É composto pelos médicos, que aguardam os exames dos pacientes; paciente, que autoriza o compartilhamento dos dados pelos provedores de saúde; e pesquisador, que solicita acesso aos estudos e aguarda os provedores de saúde enviarem. Também é descrita a comunicação dos provedores de saúde (que armazenam as imagens) com o *framework* proposto, com o objetivo de prover privacidade aos pacientes. Ademais, no campo em vermelho a figura apresenta um cenário de ataque de re-identificação, em que um usuário malicioso tenta inferir a identidade dos pacientes a partir de dados obtidos dos médicos e pesquisadores. Por fim, a partir da proposta deste projeto para o médico a privacidade é protegida pelo método K-Anonimato e para o pesquisador com a Privacidade Diferencial.

Para avaliar o *framework* frente aos ataques de re-identificação como ilustra a Figura 7,

os quadros em vermelho indicam um usuário maliciosos que visa recuperar dados dos médicos e pesquisadores para inferir em uma base pública. O cenário de ataques desta pesquisa exhibe um ataque de ligação de registros em que há uma base de dados que contém *quasi-identificadores* dos pacientes, que são: gênero, código postal e idade. Os dados disponibilizados estão em uma base pública em um serviço de saúde arbitrário, podendo ser até mesmo uma rede social. Primeiramente, o usuário malicioso tenta recuperar dados do pesquisador, como ilustra a parte superior da Figura 7, uma vez que, ele obtém sucesso na tarefa tem como resposta um conjunto de dados agregados (utilizando a Privacidade Diferencial). Na parte inferior da Figura 7 é representado um ataque de re-identificação para o médico, visto que um usuário malicioso tenta recuperar dados, tendo como resposta um conjunto de dados anonimizados (utilizando o K-Anonimato). Por fim, o usuário malicioso tenta extrair alguma informação útil, com o propósito de descobrir a quem se refere os registros obtidos, e assim fazer a ligação com a base de dados públicos.

A descrição do cenário de ataques é importante para criar uma visão das possíveis ameaças ao sistema, bem como testar a efetividade dos métodos de anonimização propostos. Desse modo, com a definição do cenário de ataques de re-identificação, fundamentado na ligação de registros com uma base pública, foi possível coletar as métricas e validar a aplicação da Privacidade Diferencial e do K-Anonimato frente ao *framework* proposto.

6.3 Avaliação de desempenho

A avaliação de desempenho do *framework* foi realizada com base em duas perspectivas, a da blockchain e a do provedor de saúde, mensurando o uso de hardware e de rede. Dessa forma, para validar a efetividade de utilizar tais métricas, foram levantados trabalhos na literatura, tais como o de Nasir *et al.* (2018), Tanwar, Parekh e Evans (2020) e Chukwu e Garg (2020) que também as empregam. Para a simulação, cada rodada de experimentos foi executada 30 vezes, com 5 rodadas – os gráficos e tabelas descritos a seguir expressam os valores das simulações.

6.3.1 Avaliação da blockchain

A avaliação de desempenho da blockchain tem finalidade de analisar a capacidade das máquinas no emprego dessa tecnologia, com o intuito de estendê-las e serem aplicadas em um ambiente real de saúde, portanto, o Hyperledger Caliper auxiliou nesta finalidade. Com essa ferramenta é possível mensurar a blockchain por completo, bem como suas partes individualmente. As métricas levantadas foram: latência, vazão, Transações por Segundo (TPS), uso de CPU e uso de memória. Para realizar as simulações os valores de TPS foram variados em: 50, 100, 150, 200 e 250 (NASIR *et al.*, 2018; TANWAR; PAREKH; EVANS, 2020).

Primeiramente, foi avaliado o consumo de hardware das máquinas virtuais na execução dos experimentos, considerando uso de CPU e uso de memória. As simulações realizadas com o

auxílio do Hyperledger Caliper foram para adição e consulta de registros na rede blockchain, apoiado pelos métodos implementados nos contratos inteligentes. Também foi analisada a execução dos métodos para cada organização que compõe a rede, são elas: provedor de saúde, pesquisador e paciente. As organizações têm o objetivo de separar as operações que cada agente pode realizar, são as seguintes: operações de adição ou consulta de ativos. A Figura 8 descreve o uso de CPU médio de cada operação durante as execuções das simulações, em que o eixo x representa as transações por segundo (TPS) e o eixo y descreve a percentagem de uso de CPU. Ademais, cada barra significa uma organização, que são agrupadas e refletem as variações de TPS citadas anteriormente. Para a operação de adição, representada na Figura 8(a), os resultados revelam que a organização do provedor de saúde é a que consome mais CPU, principalmente quando tem o TPS próximo a 200. No entanto, quando é enviado um TPS de 250 ou mais o uso de CPU diminui, uma vez que, a blockchain rejeita as transações, por ser um fluxo muito alto e é possível que seja um ataque para causar indisponibilidade no sistema. A Figura 8(b) representa a utilização de CPU para a consulta dos dados e os resultados expressam menor uso de CPU do que a adição, com a organização mais custosa sendo a do paciente, tendo em vista que os pacientes interagem constantemente com esse método para recuperar solicitações de acesso aos seus dados e aprová-las.

Figura 8 – Resultados das simulações para a quantidade de CPU em percentagem que a blockchain consome por organização com diferentes valores de transações por segundo (TPS).

Para complementar os resultados das avaliações de desempenho da blockchain foi mensurado o uso de memória pelos componentes da estrutura. A Figura 9 ilustra os resultados que descrevem as operações de adição e consulta para cada organização, representadas pelas barras. Além disso, a avaliação foi caracterizada pelo eixo x sendo as variações TPS e o eixo y o uso de memória em Megabytes (MB). A Figura 9(a) descreve o uso de memória por organização, na operação de adição dos dados, que reflete o provedor de saúde como organização que consome mais memória, uma vez que, essa organização faz paralelo com o componente do provedor de

saúde do *framework* e se comunica com o observador para adicionar dados dos pacientes na blockchain e enviá-los para anonimizar. A Figura 9(b) retrata o uso de memória para a operação de consulta, que relata o provedor de saúde como maior consumidor de memória, isso pelo fato, dessa organização recuperar os metadados, enviá-los para anonimização e encaminhá-los para os pacientes ou profissionais de saúde. Também vale ressaltar que, a organização provedor de saúde tem um par que é a âncora da rede e auxilia na comunicação com o ordenador, consumindo mais recursos devido à carga recebida por auxiliar no estabelecimento da comunicação.

Figura 9 – Resultado das simulações para a quantidade de memória que a blockchain consome por organização com diferentes valores de transações por segundo (TPS).

As métricas que correspondem ao uso de CPU e memória consumidos pela blockchain relatam um comportamento anormal, pois o esperado seria o aumento do consumo de CPU ou memória deveria seguir o aumento do TPS. No entanto, as Figuras 8 e 9 não apresentam esse comportamento, uma vez que, o Hyperledger Caliper que foi utilizado para executar os experimentos pode apresentar problemas, por ser uma plataforma nova que ainda está passando por testes. Uma hipótese sobre esse problema é que algumas transações submetidas podem falhar, isso faz as transações que estão aguardando para serem contabilizadas saírem da fila, mesmo que executadas com sucesso. Esse problema até o momento (fevereiro de 2021) não foi resolvido e ainda está em aberto no Github do Caliper e corresponde ao problema #736¹.

O segundo passo para avaliar o desempenho da blockchain no cenário proposto é a carga gerada na rede, que foi mensurada com base nas métricas latência e vazão (*Throughput*), também utilizando a ferramenta Hyperledger Caliper. A Figura 10 ilustra os resultados das simulações para cada uma das métricas, nas operações de adição e consulta dos dados, que são representadas pelas barras nos gráficos. A latência geral é descrita pela Figura 10(a), com valores próximos a 100 segundos para a operação de adição e, para a consulta, com valores próximos de 25 segundos. Essa carga maior para adição e consulta é gerada devido à aplicação do processo de

¹ <https://github.com/hyperledger/caliper/issues/736>

consenso para a validação das transações (BACH; MISAKJEVIC; ZAGAR, 2018), contudo, a carga gerada ainda é menor quando comparada com um algoritmo baseado em PoW. Com relação à vazão representada na Figura 10(b), para a operação de adição esta apresenta um fluxo máximo próximo de 4 Kilobits por segundo (Kb/s), já para a consulta está próximo de 15 Kb/s. A operação de adição gera maior latência e menor vazão para rede, visto que esse processo exige um acordo entre os pares para persistir o dado, além de o cliente ter que enviar um *JavaScript Object Notation* (JSON) com 128 KB para o par âncora, que se comunica com o ordenador, para, assim, adicionar os dados ao banco, gerando um custo maior.

Figura 10 – Resultado das simulações para o desempenho da rede para as operações de adição e consulta, considerando as métricas de latência e vazão.

Em suma, as métricas de latência e vazão apresentam um comportamento anormal. Primeiramente, é esperado que os valores de vazão fossem diretamente proporcionais aos valores de TPS, contudo, os valores variam, por exemplo, para a consulta a maior vazão tem TPS igual a 200. Nessa operação a vazão e a latência apresentam valores anômalos, não seguem o aumento do TPS. Essas adversidades são causados por um problema no Hyperledger Caliper (ver o problema #512²) que quando os valores de TPS aumentam muito, erros são causados no processamento das transações. Outro problema causado pelo Caliper e Fabric está relacionado com a definição das políticas de endosso, pois impactam no desempenho da rede (segundo o problema #321³ do Caliper). Essas políticas definem quais nós executam as transações, e devem ser projetadas com as regras de negócios para não causarem gargalo na rede. Portanto, para esta pesquisa não consideramos o projeto das políticas, assim, é gerado um gargalo maior nos experimentos.

² <https://github.com/hyperledger/caliper/issues/512>

³ <https://github.com/hyperledger/caliper/issues/321>

6.3.2 Avaliação do módulo do provedor de saúde e de privacidade

Como foi descrito no Capítulo 5 o *framework* proposto tem um módulo chamado provedor de saúde, que se complementa com o módulo de privacidade, além de enviar e recuperar dados da blockchain. Assim, a comunicação desses módulos acontece na rede do provedor de saúde que recupera dados de suas bases e os enviam para anonimização. Desse modo, com a finalidade de avaliar a carga que os módulos geram aos sistemas de saúde, foram executadas simulações a partir da ferramenta JMeter⁴, com o propósito de coletar as métricas de desempenho. As métricas analisadas foram latência e vazão para os algoritmos, tal como os Bytes enviados, configurado no cenário de aplicação descrito anteriormente e executados 30 vezes com 5 rodadas. As simulações para o compartilhamento das imagens médicas foram executadas seguindo os métodos: (i) não aplicando privacidade; (ii) aplicando privacidade com o K-Anonimato; e (iii) com a Privacidade Diferencial. Portanto, as métricas presentes na Tabela 5 são relativas aos valores médios da comunicação geral entre o módulo do provedor de saúde e de privacidade, que utilizam de operações de adição na blockchain e de consulta na base de dados do provedor de saúde.

Tabela 5 – Desempenho da rede quando se aplicam os módulos do provedor de saúde e de privacidade em conjunto com a blockchain. Foram consideradas as métricas, latência, vazão, *bytes* enviados e tempo, no qual as duas primeiras foram para avaliar o custo geral da rede comunicando com a blockchain e, a última, o tempo, foi para verificar o custo de tempo de anonimização para cada método de privacidade.

	Latência (s)	Vazão (Kb/s)	Bytes enviados	Tempo para anonimizar (ms)
Sem privacidade	1,51	5,50	284,00	Não se aplica
K-anonimato	4,74	0,14	126,00	66,05
Privacidade diferencial	201,19	0,04	126,00	31,84

Com os módulos sendo executados localmente e a blockchain em uma máquina virtual na GCP, um atraso considerável é gerado, devido ao processo de consenso na blockchain, como pode ser observado na Figura 10. Quando os modelos de privacidade são utilizados, o custo para a rede é consideravelmente maior, pois esse processo é realizado com algoritmos separados, que tomam como base um conjunto de amostras para aplicarem a privacidade. Para o K-Anonimato combinado com a blockchain, a latência na rede é de 45,47 segundos, a vazão de 0,14 Kb/s e 126 *bytes* são enviados por transação – esse método provê privacidade média aos sistemas, protegendo-os de ataques de ligação aos registros (BARBOSA; BRITO; ALMEIDA, 2020). A Privacidade Diferencial apresenta um método com alta entropia, protegendo os registros dos pacientes de ataques de ligação a bases similares e probabilísticas. Com relação ao custo para a rede, este tem alta latência como 201,19 segundos e baixa vazão de 0,04. Esse método tem um custo maior para a rede, pois precisa recuperar um número significativo de amostras

⁴ <https://jmeter.apache.org/>

para gerar o ruído e calcular a sensibilidade, dessa forma, ele gera igualmente várias trocas de mensagens entre os elementos da rede. O custo de tempo para os algoritmos isolados apresentam os respectivos tempos para anonimizar: 66,05 milissegundos (ms) para o K-Anonimato; e 31,05 ms para a Privacidade Diferencial.

Vale destacar que, na Tabela 5 contém alguns valores discrepantes para a latência da Privacidade Diferencial e o tempo de anonimização para o K-Anonimato. Inicialmente, a Privacidade Diferencial realiza diversas consultas nas bases de dados que estão distribuídas, para calcular a sensibilidade dos dados utilizados, esse processo é realizado uma única vez. Desse modo, ocorrem atrasos para iniciar o algoritmo e obter o valor do Δf (sensibilidade), pois necessita fazer uma quantidade extensiva de consultas. No entanto, o passo final de aplicar o ruído e fazer a consulta estatística é rápido, assim, tem o tempo para anonimizar menor, pois necessita somente de uma operação. O tempo para anonimizar no K-Anonimato é maior que na Privacidade Diferencial, tendo em vista que a iniciação do algoritmo é rápida, pois somente manipula constantes. Porém, o passo final faz a manipulação de vários atributos para ocultá-los a partir de expressão regular ou generalizá-los em classes, assim, o tempo para anonimizar é maior para o K-Anonimato.

6.4 Avaliação dos requisitos de privacidade

Esta seção apresenta o processo de avaliação da privacidade, e seus resultados. Com base no cenário de aplicação e nas questões envolvendo ataques de ligação, foram executadas simulações e coletados os valores anonimizados para os modelos K-Anonimato e Privacidade Diferencial serem avaliados. Além disso, foram analisados os requisitos que o *framework* pode proporcionar aos sistemas de saúde, com as metodologias da OWASP e PbD. A seguir a avaliação é descrita com mais detalhes.

6.4.1 Avaliação qualitativa da privacidade

Para esta avaliação, foi tomada como foco a variável condição do sistema – tal variável foi examinada a partir dos requisitos definidos pela OWASP, que é uma comunidade *Open Source* para a definição de metodologias, ferramentas e levantamento dos riscos aos quais estão sujeitas as aplicações convencionais que compõem a Internet (OWASP, 2017). Assim, a OWASP criou uma lista com os 10 riscos comuns à privacidade para as aplicações que estão na Internet, fazendo paralelo, igualmente, com os predicados da metodologia *Privacy by Design* (PbD) (BARBOSA; BRITO; ALMEIDA, 2020). Com o levantamento desses riscos é discutido como o *framework* proposto auxilia a mitigar cada um deles, focando na **variável condição do sistema**. São eles:

1. **Vulnerabilidade da aplicação:** está associada à API de comunicação com os clientes, podendo apresentar vulnerabilidades e estar suscetível a ataques, por exemplo, ataques

de negação de serviço. No entanto, a exploração das vulnerabilidades da API não está no escopo desta pesquisa, assim, podem apresentar outros problemas clássicos ou até desconhecidos. O *framework* distribuiu todas as operações com a replicação dos serviços, ou seja, se algum nó gerenciador cair, outra cópia confiável toma conta como nó principal;

2. **Vazamento de dados do lado do operador:** os provedores de saúde realizam o tratamento dos dados pessoais dos pacientes com seu consentimento. Para evitar o acesso indevido aos dados na blockchain, o HFL tem um esquema de criptografia de Curvas Elípticas, com a curva prime256v1 e resumos *hash* SHA-256 (FABRIC, 2020b). Além disso, o *framework* dispõe de um esquema de *token* único, que é associado às imagens recuperadas do banco de dados, para auditar e localizar vazamentos. Por fim, as imagens que são compartilhadas ficam distribuídas a partir da rede IPFS, desse modo, para acessá-las é necessário ter conhecimento da chave de acesso, que é um *hash*;
3. **Resposta insuficiente ao vazamento de dados:** o *framework* também foi desenhado com a finalidade de proteger os sistemas de saúde contra vazamentos, mas, se ocorrerem, é possível responsabilizar a organização que não fez o tratamento adequado, com base nos logs imutáveis na blockchain e o *token* de rastreamento na imagem. Em suma, os logs armazenados na blockchain são imutáveis, logo, se uma organização acessar um dado para pesquisa, os logs são armazenados e livres de adulteração. Essa característica é atendida parcialmente, pois, os logs auxiliam na resposta a vazamentos, mas, não foi considerada uma análise reativa de incidentes e nem medidas para a previsão de vulnerabilidade;
4. **Exclusão insuficiente de dados pessoais:** para a exclusão dos dados pessoais o processo não é tão simples e tem que passar por análises através de pares confiáveis, que são os provedores de saúde e os pacientes que o ativo pertence. Também para a exclusão da imagem, é necessário remover o vínculo com a rede do IPFS, no entanto, os logs de acesso ainda ficam disponíveis na blockchain. Para complementar essa questão deve ser observado na LGPD, pois, a lei trata do direito a exclusão dos dados. Esse requisito é atingido parcialmente, devido aos logs no blockchain serem imutáveis;
5. **Políticas, termos e condições não transparentes:** qualquer acesso aos dados do paciente devem ser permitidos por ele, seguindo as políticas e normas estabelecidas pela LGPD. Esse acordo também foi estabelecido fundamentado em contratos inteligentes, fornecidos pela blockchain;
6. **Coleta de dados não necessária para o objetivo principal:** devido ao fato do *framework* ser projetado para um objetivo específico, que é o compartilhamento de dados de saúde, a coleta também é específica. Assim, todos os dados coletados que são requisitados por um pesquisador ou médico necessitam do consentimento do paciente para serem utilizados;
7. **Compartilhamento de dados com terceiros:** por utilizar um *framework* blockchain e P2P como a rede IPFS, os dados não passam por um terceiro confiável. Os dados são tratados

de forma descentralizada, sem ponto único de falha para acessá-los. Para acessarem os dados o perfil de acesso é legítimo, que restringe que os usuários apenas acessem pelas organizações e por canais definidos, logo, somente médicos acessam os dados em seu nome;

8. **Dados pessoais desatualizados:** esse ponto é irrelevante para o *framework*, pois mesmo dados antigos são importantes para a pesquisa, sendo assim, serão úteis. No entanto, se os dados atualizados forem importantes para os provedores de saúde manterem as condições dos pacientes atualizados, é possível enviar uma solicitação para atualização do dado na blockchain;
9. **Vencimento de sessão ausente ou insuficiente:** para a construção da blockchain no *framework*, foi considerada a plataforma HLF, que gerencia as sessões na rede e controla o fluxo das transações entre os usuários e a blockchain. Além disso, o tempo para o estabelecimento de uma sessão pode ser configurado no perfil de conexão quando a rede é gerada, em que para esta proposta foi utilizado o padrão do HFL que são 100 segundos (FABRIC, 2020b);
10. **Transferência insegura de dados:** o *framework* tem como um dos objetivos principais prover a transferência segura de dados. Desse modo, se utiliza de algoritmos de anonimização e a transferência de dados com *Transport Security Layer* (TSL) através do HFL para assegurar a transferência entre cada nó e organização (FABRIC, 2020f).

Em síntese, com base nesta avaliação e na metodologia da OWASP, foi possível observar que os 10 requisitos listados são alcançados por completo ou parcialmente pelo *framework* proposto. Nesse sentido, é possível ter um sistema seguro, que garante a privacidade do paciente e que segue as políticas de privacidade expressas pela LGPD. Ademais, o *framework* segue os conceitos da PbD que, por padrão asseguram a privacidade e salvaguarda de problemas com a privacidade antes que estes aconteçam. Por fim, a privacidade que o *framework* proporciona é centrada nos pacientes (titulares dos dados), que são os agentes afetados quando ocorre vazamento de dados.

6.4.2 Avaliação quantitativa da privacidade

Esta seção analisa os resultados dos métodos de privacidade aplicados a blockchain para anonimizar os dados. Em vista disso, pode-se medir a resistência dos métodos a ataques de ligação ou probabilísticos. Os experimentos foram executados em uma máquina MacOS Catalina 13.10, com 8 GB de RAM, processador Intel Core i5 e armazenamento *flash* de 256 GB. Cada rodada de experimento foi executada 30 vezes, com 5 rodadas. Segundo Liu *et al.* (2018) o intervalo padrão para a técnica de Privacidade Diferencial, está entre 0 e 1. Para cada um dos modelos avaliados, as métricas extraídas foram entropia e utilidade que, segundo Wagner e Eckhoff (2018) são relevantes para examinar a privacidade de um sistema.

A princípio, considerando os modelos de privacidade empregados, a Privacidade Diferencial em específico apresenta uma questão a ser analisada, que é o nível de ruído, dado pelo símbolo ϵ na fórmula que descreve o modelo. Assim, é necessário verificar qual é o valor de ϵ adequado para que seja efetiva na aplicação da privacidade sobre as tabelas de dados. A Figura 11 apresenta o estudo realizado para os valores do ϵ no intervalo de 0 até 1, comparando o nível de privacidade, representado pela entropia, assim como a utilidade dos dados, representado pela dissimilaridade. Portanto, os valores no gráfico refletem que, conforme aumentam os valores do ϵ , cresce a quantidade de entropia até esta ficar linear. Dessa forma, a proporção que aumenta o ϵ , diminui a utilidade dos dados. Afinal, para atingir um nível adequado de privacidade, tal como de utilidade de dados, o ϵ escolhido foi igual a 0,01, pois, na Figura 11 é o ponto de encontro entre a entropia e utilidade, ou seja, é quando essas métricas atingem um ponto médio, com dados ainda úteis, mas também privados.

Figura 11 – Influência da escolha do ϵ para as métricas entropia e utilidade. A métrica entropia é representada pela curva laranja e a utilidade pela curva verde, que são descritas seguindo a quantidade (eixo y) pelo valor do ϵ (eixo x).

A definição do ϵ igual a 0,01 para todas as simulações, o passo seguinte foi avaliar a quantidade de privacidade que os modelos garantem, seguindo a métrica de entropia, que mede a desordem dos símbolos anonimizados (WAGNER; ECKHOFF, 2018). Assim, quanto maior o valor da entropia, mais difícil é de um nó malicioso inferir o valor real a partir do anonimizado (WAGNER; ECKHOFF, 2018). A Figura 12 descreve o *boxplot* para os métodos K-Anonimato e Privacidade Diferencial, em que no eixo y representa a quantidade de entropia normalizada e o eixo x os modelos de privacidade. Desse modo, o K-Anonimato tem a mediana da entropia próxima de 0,5, com variância de 0,04 e desvio padrão de 0,21. Já para a Privacidade Diferencial,

a mediana é de 0,72, com variância de 0,07 e desvio padrão de 0,27, portanto, as quantidades de entropia convergem para valores elevados. Isso implica uma maior variabilidade de entropia para os valores dos atributos. Comparando a entropia para os dois modelos, pode-se considerar que a Privacidade diferencial apresenta valores mais distantes da média, logo, esse modelo tem maior variabilidade, refletindo em menor probabilidade de um usuário malicioso extrair algum conhecimento dos dados anonimizados para descobrir o valor real.

Figura 12 – *Boxplot* para as quantidade de entropia de cada modelo de privacidade. Assim, o eixo x descreve os modelos utilizados, o K-Anonimato e Privacidade Diferencial e o eixo y apresenta a quantidade de entropia $H(X)$ normalizada.

As questões de pesquisa **QP** (*Como mitigar os riscos à privacidade dos pacientes no compartilhamento de dados de saúde em redes baseadas em blockchain?*) e **QS 2** (*Como mitigar ataques de vinculação para prover privacidade aos pacientes em ambientes de saúde quando compartilham seus dados?*) são verificadas com base na entropia, descrevendo que os ataques de re-identificação são menos prováveis quando aplicados ambos os modelos de privacidade, sendo mais efetivo quando aplicado a Privacidade Diferencial, pois tem maior entropia, mas menos utilidade (WAGNER; ECKHOFF, 2018). Ao se aplicar o modelo de Privacidade Diferencial, um nível maior de privacidade é obtido, pois, a entropia converge para valores elevados. Portanto, esses resultados expressam que para mitigar ataques que fazem a ligação de tabelas similares ou probabilísticas, podem ser aplicados os dois mecanismos, uma vez que o K-Anonimato tem menor efetividade, mas auxilia na redução. Esse mecanismo pode ter apresentado menor efetividade devido ao valor de K utilizado, que deve ser estudado em busca de um valor que resulte em maior entropia. Por sua vez, a Privacidade Diferencial aplicada ao blockchain como meio para prover privacidade apresenta uma entropia elevada, respondendo à questão **QS 2**

por completo e parte da **QP**. A respeito da questão principal, cabe ressaltar que os próximos resultados auxiliam na resolução dela.

Para avaliação do comportamento dos métodos de privacidade, é calculada a utilidade, que é uma métrica que avalia a semelhança de um dado anonimizado público com seu valor real (WAGNER; ECKHOFF, 2018). Essa métrica representa a utilidade dos dados para que estes possam ser aproveitados pelo usuário final, com o objetivo de extrair alguma informação. Nesse cálculo, foi empregado a fórmula da distância de cossenos, apresentada na Seção 3.6, para distância entre os atributos anonimizados e reais. A Figura 13 representa o *boxplot* das distâncias dos dois modelos de privacidade, comparando o valor real com o anonimizado, no qual o eixo y refere-se à utilidade normalizada para cada modelo, e o eixo x representa os modelos que foram analisados, o K-Anonimato e a Privacidade Diferencial. O K-Anonimato tem mediana da utilidade igual a 0,26, variância de 0,0008 e desvio padrão de 0,36, assim, esses valores indicam a menor distância entre os atributos reais e o anonimizado, tal como menor variabilidade, convergindo para próximos à média. Para a Privacidade diferencial, a utilidade é maior, com mediana igual a 0,49, variância de 0,010 e desvio padrão de 0,103, portanto, a utilidade apresenta valores mais afastados da média e também valores centrais maiores. Dessa forma, o modelo do K-Anonimato apresenta melhor utilização dos dados em relação à Privacidade Diferencial, assim, com o K-Anonimato é possível o usuário final ter privacidade e, seus dados ainda serem representativos para pesquisa. Consequentemente, os resultados são complementares ao cálculo da entropia, afinal, a privacidade segue paralelamente à utilidade, pois quanto maior o nível de utilidade, menor a privacidade.

Análise do comportamento das grandezas entropia e utilidade

Com o objetivo de investigar a caracterização das métricas entropia e utilidade (dada pela distância) para ambos os modelos de privacidade, estes foram comparados a partir das funções de densidade de probabilidade, do inglês, *Probability Density Function* (PDF), que analisa a densidade dos valores de entropia e a utilidade das amostras coletadas. Além disso, foi empregada a função densidade de probabilidade acumulada, do inglês, *Cumulative Density Function* (CDF), para estimar os parâmetros, com base em algumas funções de densidade comuns, de modo a definir as que mais se aproximam da curva real. As distribuições utilizadas foram: *Gamma*, *Normal*, *Qui-quadrada*, *Exponencial*, *T-student* e *Log-normal* (JAIN, 1991). Vale ressaltar que, dentre as distribuições estudadas, em cada gráfico foram adicionadas somente as três distribuições que melhor se ajustaram com os valores reais, para a entropia e utilidade, respectivamente. Logo, para calcular o ajuste foi empregado o teste *Komogorov-Smirnov* (K-S). Todos os testes estatísticos foram realizados com a linguagem R e o *R Studio*.

A Figura 14 apresenta as PDFs das métricas para as amostras coletadas dos experimentos, em que o eixo x representa as distâncias, ou seja, a utilidade e o eixo y a densidade dos valores da utilidade. O intervalo de confiança utilizado na definição das distribuições foi de 95% e são indicados pelas linhas pontilhadas, sendo as linhas em vermelho para o K-anonimato e em

Figura 13 – *Boxplot* para a utilidade dos dados extraído por cada modelo de privacidade, dado pela fórmula 1 menos a distância de cossenos. Assim, o eixo x descreve os modelos utilizados, o K-Anonimato e Privacidade Diferencial e o eixo y apresenta a dissimilaridade normalizadas.

azul para a Privacidade diferencial. As distribuições expressas na Figura 14(a) apontam que, para o K-anonimato, os valores de distância entre os atributos têm picos entre 0,3 e 0,4, já para a Privacidade diferencial tem picos entre 0,4 e 0,5. Portanto, o K-anonimato tem maior concentração para a direita, apresentando, assim, menor distância entre os atributos, o que implica que os dados após anonimizados com esse modelo têm maior utilidade para o usuário final. Já a PDF da entropia é descrita pela Figura 14(b), que representa os dois modelos de privacidade. O K-anonimato tem picos próximos de 0,5 e uma distribuição com tendência para o centro do gráfico, assim, ele pode apresentar valores médios de privacidade aos atributos, quando empregados. A Privacidade diferencial tem picos próximos de 0,75 e maior concentração direcionada para a esquerda, ou seja, esse modelo apresenta maior privacidade comparado com o outro modelo apresentado. Portanto, pode-se observar estatisticamente que a Privacidade diferencial fornece maior privacidade, mesmo quando é apresentada uma generalização dos valores com uma PDF.

Outro método utilizado para explorar os resultados das métricas entropia e utilidade foi a CDF para cada uma delas, que serve como uma forma de caracterizar o comportamento das métricas e até mesmo de prever novos valores. Desse modo, é possível generalizar o alcance das métricas e a eficiência dos métodos de privacidade (JAIN, 1991). Logo, com o objetivo de descrever a tendência do comportamento das variáveis entropia e utilidade, foram empregadas as CDFs comuns na literatura, como discutido anteriormente, para, assim, validar a proximidade entre as funções, comparadas com a função que descreve os valores reais. Para comparar

Figura 14 – Caracterização das métricas de privacidade para a função densidade de probabilidades. O eixo x para a Figura 14(a) representa as distâncias com base na fórmula da dissimilaridade. Para a Figura 14(b) o eixo x representa a entropia, isto é, o nível de privacidade. O eixo y para ambas as figuras representam a densidade, ou seja, a probabilidade de concentração dos valores.

numericamente a proximidade entre as funções teóricas e a empírica para a variável entropia é exposta na Tabela 6 com os resultados comparativos no teste K-S.

Tabela 6 – Proximidade das funções densidade de probabilidade acumulativa para os valores reais de entropia a partir do teste *Komogorov-Smirnov* (K-S).

	CDFs para a entropia					
	Gamma	Normal	Qui-quadrado	Exponencial	T-student	Log-normal
K-anonimato	0,70	0,46	0,53	0,46	0,50	0,63
Privacidade diferencial	0,55	0,66	0,66	0,33	0,66	0,55

Após o cálculo do ajuste das curvas, foi representado graficamente, na Figura 15, as CDFs da entropia para o K-Anonimato e Privacidade Diferencial. Como pode-se observar na Tabela 6, para o modelo K-Anonimato as funções que melhor se ajustam, ou seja que tem a distância máxima menor para a curva empírica são: Exponencial, Normal e *T-student*. A Figura 15(a) descreve o comportamento das distribuições citadas para o K-Anonimato, em que as funções são representadas por diferentes cores, o eixo y descreve a probabilidade acumulada e o eixo x a entropia. Assim, as distribuições que melhor se ajustaram é a exponencial e a normal. Desse modo, podemos concluir que a variável entropia, tem comportamento semelhante à função densidade de probabilidade Exponencial ou Normal. Já para a Privacidade Diferencial, representada pela Figura 15(b), as funções que melhor se ajustaram foram: *Gamma*, Exponencial e *Log-normal*. Portanto, entre essas funções a que melhor se ajustou foi a Exponencial, que possui o menor valor para o teste K-S. Em suma, o comportamento da variável entropia para o modelo de Privacidade Diferencial tem comportamento próximo de uma função Exponencial.

Figura 15 – Caracterização das funções densidade de probabilidade acumulada CDFs para a variável entropia, com representação para cada método de privacidade, K-anonimato e Privacidade diferencial.

Para a variável que representa a utilidade dos dados, dada pela distância entre os atributos reais e os anonimizados, também foi analisado o comportamento a partir das CDFs, comparando as curvas teóricas frente à curva empírica. Assim, também é necessário explorar as curvas que melhor se ajustam com a curva empírica. Nesse sentido, a Tabela 7 representa o teste K-S para as funções que foram extraídas da literatura. Dessa forma, os ajustes comparados foram representados graficamente, com base nas três curvas que melhor se ajustaram. A seguir, são apresentados os gráficos para a CDFs da utilidade para ambos os modelos de privacidade.

Tabela 7 – Proximidade das funções densidade de probabilidade acumulativa para os valores reais da utilidade, a partir do teste *Komogorov-Smirnov* (K-S).

	CDFs para a utilidade					
	Gamma	Normal	Qui-quadrado	Exponencial	T-student	Log-normal
K-anonimato	0,96	0,70	0,80	0,60	0,56	0,83
Privacidade diferencial	0,85	0,57	0,53	0,54	0,63	0,60

Com os ajustes calculados, a distância máxima entre as funções que apresentaram menor distância (max-min) foram calculadas, com base no teste K-S, com propósito de calcular as distâncias das curvas teóricas para a empírica. A Figura 16 apresenta as CDFs para os dois modelos de privacidade, o eixo x representa a distância, ou seja, a utilidade, enquanto o eixo y descreve a probabilidade acumulada. Dessa maneira, a Figura 16(a) descreve o comportamento das CDFs que melhor se ajustam para o modelo K-Anonimato. As selecionadas foram: Exponencial, Normal e T-student. Considerando essas funções, a que mais se aproxima é a T-student, ou seja, a variável utilidade para o K-Anonimato tem comportamento similar a uma função T-student. Para o modelo Privacidade Diferencial, expresso na Figura 16(b), também foram

selecionadas as funções que melhor de adaptam aos valores reais, as quais foram: Exponencial, Normal e Qui-quadrada. Seguindo essas CDFs, a que melhor se adapta à variável utilidade é a Qui-quadrada, assim, o comportamento descrito pela utilidade pode ser similar ao de uma função Qui-quadrada. Observando o comportamento dessa métrica, fundamentada pelas CDFs, é possível prever o comportamento para outros valores, provando, desse modo, que a variável converge para valores menores para o K-Anonimato e maiores para a Privacidade Diferencial.

Figura 16 – Caracterização das funções densidade de probabilidade acumulada (CDFs), para a variável utilidade, com representação para cada método de privacidade, K-anonimato e Privacidade diferencial.

Por fim, as curvas apresentadas nos resultados para a PDF e CDF podem ser utilizadas como estimadores para os parâmetros, entropia e utilidade, com confiabilidade de 95%, a partir dos intervalos de confiança. Nesse sentido, podem ser utilizadas para descrever o comportamento futuro desses parâmetros, com base em curvas teóricas estabelecidas na literatura e provando a fidelidade dos experimentos. Portanto, com a observação do comportamento das métricas, é possível compreender que, em outros casos, ou seja, com outros parâmetros, torna-se viável atingir um valor confiável de entropia e utilidade adequado, como foi descrito nos experimentos.

6.5 Análise teórica do *token* de rastreabilidade

No que concerne o levantamento sobre o rastreamento de ativos compartilhados e responsabilização de organizações pelo tratamento inadequado de dados, além de seguir as políticas da LGPD, é necessário analisar como o *token* se comporta. Para esse fim, foi empregado a análise do *hash* gerado para o *token* de rastreamento e a probabilidade de resistir a ataques de colisão (do inglês, *Collision attacks*). Tendo em vista o problema da colisão de *hashes*, foi realizada a análise teórica fundamentada no ataque do aniversário, para tentar invalidar os *hashes* ou adulterá-los (SOLTANIAN; AMIRI, 2016). Em suma, para avaliar o ataque do aniversário

– é um ataque que surge do paradoxo do aniversário, que infere qual a probabilidade de uma sala com n pessoas terem a mesma data de aniversário. Isso é análogo a um *hash*, ou seja, qual a probabilidade de encontrar um *hash* similar no conjunto de *tokens* (STALLINGS, 2014) –, é necessário calcular a probabilidade de colisão, usando números de bits (com o *Secure Hash Algorithms* SHA-256) para a contagem dos valores de colisão (SU; XIE; Lü, 2016).

De acordo com Al-Ani *et al.* (2019) a probabilidade de colisão pode ser calculada utilizando a Equação 6.1, descrevendo a quantidade de mensagens enviadas e a Equação 6.2, que é a probabilidade de colisão. Primeiramente, é calculada a quantidade de mensagens enviadas, isto é, N_m , em que BW é largura de banda, MZ o tamanho da mensagem enviada e T o tempo de envio. Com o valor do N_m coletado, foi calculada a probabilidade de colisão P , com base na equação 6.2, em que n é número de bits da mensagem. Para extrair as métricas a serem aplicadas nas equações, foi utilizada a ferramenta *Wireshark*⁵, resultando nos seguintes valores: $BW = 9,2$ Mbytes, $MZ = 671$ bytes, $T = 10$ s, e $n = 256$. Os valores foram convertidos para bytes, como pode ser observado nas equações.

$$N_m = \frac{BW}{MZ} * T \quad (6.1)$$

$$N_m = \frac{9.2 * 10^{20}}{671} = 1.37 * 10^{19}$$

$$P = \frac{N_m}{2^n} \quad (6.2)$$

$$P = \frac{1.37 * 10^{19}}{2^{256}} = 1.18 * 10^{-58} \%$$

Dessa forma, o resultado teórico alicerçado nas equações (seguindo o método de Al-Ani *et al.* (2019)) para a probabilidade de colisão foi de $1,8 \times 10^{-58} \%$, que representa a possibilidade de um usuário malicioso encontrar um *hash* similar no conjunto de *tokens*. Assim, para um nó malicioso realizar um ataque de força bruta ou de aniversário, que seja eficiente para fraudar o *token*, o custo computacional é elevado, devido ao tamanho da palavra e de sua aleatoriedade. Além de serem resistentes a ataques de colisão, as funções *hash* devem ser resistentes inversão, visto que é computacionalmente difícil encontrar um valor arbitrário X , que resulta em um *hash* conhecido (STALLINGS, 2014). Contudo, para definir o modelo de ataques a *hash* foi tomado como base trabalhos da literatura como em Al-Ani *et al.* (2019), que abordam ataques de colisão. Para este trabalho não foram considerados ataques que verificam a resistência a reversão, somente ataques de colisão. Métodos de ataques de resistência podem ser adicionados em um trabalho futuro.

⁵ <https://www.wireshark.org/>

6.6 Discussão

Esta seção apresenta os principais resultados, considerando os experimentos e dados descritos anteriormente. Dessa forma, são apresentados os resultados para desempenho da blockchain e do módulo do provedor de saúde e quais foram suas influências aos sistemas de saúde. Também são discutidos os resultados para as métricas de entropia e utilidade para cada um dos modelos de privacidade empregados, que auxiliaram na resolução da questão de pesquisa principal **QP**. Por fim, são discutidos os pontos que esta pesquisa avança frente aos trabalhos relacionados levantados. Os resultados expressos anteriormente, tiveram o objetivo de apresentar como o *framework* reduz os problemas relativos à privacidade de identidade e das transações na blockchain para saúde. Em suma, descritos pelas métricas de entropia e utilidade, que implicam em uma nível de privacidade adequado.

Os resultados correspondentes ao desempenho foram analisados no cenário de aplicação para o compartilhamento de imagens médicas, com as seguintes métricas: uso de CPU e memória, bem como latência, vazão e bytes enviados. A adição prática da blockchain aos sistemas de saúde é aceitável, pois, em questão de hardware, foi capaz de realizar as simulações consumindo no máximo 12% de CPU e 200 MB de memória. Logo, a aplicação em um ambiente real, como a GCP, na qual foram executados os experimentos, tem implementação viável, com custo mensal de cerca de R\$735,00, segundo o medidor de custos da GCP, para as máquinas utilizadas nas simulações. Já o desempenho da rede blockchain, possui uma carga considerável para latência, uma baixa vazão e os bytes enviados são cerca de 120 bytes por transação, apresentando esses custos em virtude do processo de consenso e algumas verificações de que necessita realizar. No entanto, a carga gerada ainda é aceitável para alguns sistemas, os que não necessitam do processamento em tempo real, uma vez que, o *framework* disponibiliza melhorias na confiabilidade, auxiliado pela blockchain, e na privacidade, auxiliado pelos métodos de privacidade. O desempenho do módulo do provedor de saúde que se comunica com os modelos de privacidade, gera um custo elevado para a rede, tendo em vista que o K-Anonimato necessita de um tempo considerável para anonimizar e a Privacidade Diferencial tem alta latência. Mesmo proporcionando maior privacidade para os usuários, o *framework* tem como limitação o custo para anonimizar os dados, ambos os métodos apresentam uma desvantagem, assim é necessário realizar um estudo mais aprofundado sobre eles, com o objetivo de otimizá-los. Em um ambiente que necessita do processamento em tempo real o custo do *framework* pode ser inicialmente inviável, no entanto, é possível melhorar o desempenho dos métodos. Esse processo não foi realizado nesta pesquisa, pois está fora do escopo, o objetivo inicial foi propor um método para melhorar a privacidade, como pode ser observado pela métrica de entropia, mas o desempenho deles não foi considerado.

Com base na questão de pesquisa **QP**: “**Como mitigar os riscos à privacidade dos pacientes no compartilhamento de dados de saúde em redes baseadas em blockchain?**”, na qual a premissa foi a utilização da blockchain para prover imutabilidade, rastreabilidade e

não repúdio, tal como a integração dos modelos de privacidade para mitigar riscos a ataques de re-identificação, resultando no *framework* proposto. Para examinar o *framework*, foi definido o cenário para o compartilhamento de imagens médicas e assim coletadas as métricas. Em vista disso, foram realizados os experimentos para verificar a privacidade que o *framework* garante quando aplicado a um sistema de saúde, logo, a primeira etapa da avaliação da privacidade foi para a condição do sistema, em que os requisitos da OWASP foram alcançados por completo ou parcialmente. Essas garantias tornam os sistemas de saúde mais seguros e privados para seguir as políticas da LGPD. Após a análise dos requisitos de privacidade, foram verificadas as métricas que a quantificam, que são a entropia e utilidade para os atributos anonimizados. Com essas métricas, foram analisados cada um dos modelos de privacidade: o K-Anonimato com menor privacidade e maior utilidade dos dados para o usuário final; e a Privacidade Diferencial com maior entropia e menor utilidade dos dados. Segundo a literatura, ambos os modelos auxiliam a mitigar ataques de ligação, além da Privacidade Diferencial reduzir os riscos a ataques probabilísticos (BARBOSA; BRITO; ALMEIDA, 2020). Os efeitos desses ataques são reduzidos devido aos registros e tabelas serem anonimizados, assim, um nó malicioso consegue extrair pouca informação dos dados coletados, que não servem como base de conhecimento para inferir outra característica do usuário, com o propósito de descobrir sua identidade. Além disso, há outros ataques que o K-Anonimato está suscetível e não tem um resultado efetivo, são eles: ataque de homogeneidade e ao ataque de conhecimento de segundo plano (ver na Seção 2.3.2) (BASSO *et al.*, 2016; SWEENEY, 2002). Para esta pesquisa foram considerados somente ataques de re-identificação, logo, para trabalhos futuros é interessante explorar outros ataques como os citados acima.

A respeito dos trabalhos relacionados, segundo os resultados apresentados, esta pesquisa contribui com um método para garantir privacidade em ambientes de saúde, além de integrar a blockchain para preencher lacunas relacionadas com a segurança, como maior confiabilidade na auditoria e não repúdio. Esta proposta também apresenta uma prova de conceito que foi simulada na GCP, diferentemente de alguns trabalhos na literatura, como Guo *et al.* (2018) que apresentam somente uma proposta teórica, ou Patel (2018) que tem um modelo, mas, não testa em ambientes fora da rede local. Ademais, o *framework* proposto confere maior privacidade aos pacientes e apresenta utilidade para serem empregados nas ações dos usuários finais, mas não o bastante para não oferecer riscos à privacidade do titular dos dados. Ao contrário, trabalhos relacionados como os de Shen, Guo e Yang (2019) e Al-Omar *et al.* (2019) integram a criptografia nos atributos ou nos textos por completo, estando sujeitos a roubo ou a perda de chaves, e que encriptados tem uma entropia elevada e nenhuma utilidade para o usuário final. Nenhum dos métodos analisados na literatura abordam técnicas para mitigar problemas com ataques de re-identificação (SWEENEY, 2002; ALMADHOUN; AYDAY; ULUSOY, 2020) (que podem ser submetidos contra o K-Anonimato e a Privacidade Diferencial) em ambientes blockchain para área saúde. Por fim, diferentemente dos trabalhos apresentados na literatura, esta pesquisa realiza a análise e caracterização estatísticas dos métodos de privacidade, para definir como as métricas de entropia

e utilidade se comportam frente aos experimentos conduzidos.

CONCLUSÃO

Neste trabalho foi apresentado um *framework* baseado em blockchain para preservação da privacidade no compartilhamento de dados de saúde, o BCPSHield, em que foi abordado um estudo piloto para o compartilhamento de imagens médicas. O *framework* emprega dois modelos de privacidade, que se baseiam em de-anonimização, com o propósito de proteger os pacientes dos riscos de vazamentos, tal como de ataques que têm o objetivo de ligar os dados privados com identificadores reais, para descobrir a identidade dos pacientes. Os modelos empregados foram o K-Anonimato e a Privacidade Diferencial que, respectivamente, garantem melhor utilidade dos dados anonimizados e maior privacidade, para reduzir, igualmente, os riscos a ataques de re-identificação. Portanto, este *framework* teve como propósito responder à questão de pesquisa principal, de modo a reduzir os riscos a privacidade no compartilhamento de dados de saúde, por meio do uso da blockchain e dos modelos de privacidade.

O *framework* permite a fácil integração de outros módulos, além da modificação dos módulos já existentes por serem baseados no Docker para a construção, bem como na rede do Hyperledger Fabric que permite que os contratos inteligentes sejam customizados. Os desenvolvedores podem implementar *logs* com diferentes características para o armazenamento. Os modelos de privacidade também podem ser customizados para o *framework*, no entanto, há riscos em modificar essa característica, pois a implementação foi própria para esta abordagem e dependendo da implementação realizada, estes podem ir contra a LGPD ou até mesmo não cobrir os requisitos da OWASP. Os métodos de privacidade utilizados apresentam garantias teóricas – como a definição teórica da Privacidade Diferencial – para definir as garantias contra os ataques de re-identificação, com entropia e utilidade para serem empregados por pesquisadores, assegurando, ainda, que a privacidade dos usuários não será invalidada. Diante disso, os resultados apresentam a viabilidade do *framework* e as vantagens da utilização para os ambientes de saúde.

Na prática, o *framework* proposto pode ser integrado a um sistema real de saúde, além de formar uma rede confiável para o compartilhamento desses dados, assegurando a privacidade

dos pacientes que partilham suas informações para pesquisa. No entanto, vale ressaltar que se o sistema necessitar do processamento em tempo real, o *framework* torna-se inviável, pois o seu custo é alto. Além disso, o *framework* também oferece um ambiente com não repúdio e imutabilidade com base na blockchain, que é viável para a integração em servidores em nuvem, como foi apresentado nas simulações executadas na infraestrutura da GCP. Por fim, a proposta pode trazer benefícios aos pesquisadores, profissionais de saúde e aos pacientes se seus dados forem utilizados pela ciência para gerar novos conhecimentos.

7.1 Principais contribuições

Esta seção apresenta as principais vantagens que o *framework* BCPSHield adiciona aos sistemas de saúde baseados na blockchain, que envolvem contribuições científicas e técnicas.

As contribuições científicas são:

- Análise sistemática de trabalhos relacionados e o levantamento das problemáticas envolvidas na aplicação da blockchain em sistemas de saúde;
- A proposta do *framework* BCPSHield, que é fundamentada em um modelo para preservar a privacidade em ambientes de saúde baseados em blockchain, para a troca segura de informações sobre pesquisas médicas;
- A adição da camada de privacidade a arquitetura de referência proposta por [Yuan e Wang \(2018\)](#). Proporcionado a ligação de dados complexos na blockchain (como imagens DICOM, *logs* de sensores, áudio e outros), um *token* de rastreabilidade desses dados e modelos de anonimização para proteger a privacidade dos usuários;
- O emprego e modificação dos algoritmos de privacidade para integração no *framework* junto a blockchain de forma adaptável, com base nos modelos: K-Anonimato e Privacidade Diferencial;
- Proposta preliminar de políticas de compartilhamento e controle de acesso com base na LGPD, que se preocupam com a remoção da ligação dos dados entre a blockchain e a rede *Off-chain*;
- A reutilização dos dados estatísticos anonimizados para a utilização em pesquisa, sem revelar a identidade do paciente por re-identificação;
- O controle das informações compartilhadas utilizando um *token* de rastreamento;
- O compartilhamento de imagens médicas protegendo a privacidade dos pacientes, uma vez que, trabalhos como [Azaria et al. \(2016\)](#), [Patel \(2018\)](#), [Gan et al. \(2020\)](#), não empregam esse tipo de dados ou não se preocupam fortemente com a privacidade;

- A análise e caracterização estatística dos modelos de anonimização K-Anonimato e Privacidade Diferencial, aplicados a blockchain, tal como a integração de uma prova de conceito para compartilhamento de imagens médicas frente a ataques de re-identificação.

As contribuições técnicas são:

- Implementação do *framework*, com a linguagem GO e o Docker, com o objetivo de ser escalável e customizável;
- Elaboração de uma biblioteca para a linguagem GO, que emprega a técnica de Privacidade Diferencial (ver no Github¹);
- Elaboração da Prova de Conceito para uma rede de saúde no Hyperledger Fabric, que preserve a privacidade e disponibilize o controle de acesso para diferentes usuários.

7.2 Produção científica

- AGUIAR, E. J.; FAIÇAL, B. S.; KRISHNAMACHARI, B.; UEYAMA, J. A survey of blockchain-based strategies for healthcare. **ACM Computing Surveys**, Association for Computing Machinery, New York, NY, USA, v. 53, n. 2, mar. 2020. ISSN 0360-0300. Disponível em: <<https://doi.org/10.1145/3376915>>.

7.3 Trabalhos futuros

Com o desenvolvimento da proposta deste trabalho, surgiram ideias para avançá-la, as quais serão listadas a seguir:

- Integrar o *framework* para o compartilhamento de registros eletrônicos de saúde em um ambiente real;
- Realizar um estudo para decidir quais os valores de K são mais adequados para o K-Anonimato e do épsilon para a Privacidade Diferencial, para, assim, aumentar a precisão em anonimizar os dados;
- Implementar uma interface de usuário mais amigável, com o objetivo do usuário final utilizar da plataforma e se comunicar com os outros membros da rede;
- Modificar o algoritmo de Privacidade diferencial para fornecer dados além dos agregados, com menor granularidade e sem sofrer riscos à privacidade do paciente;

¹ <https://github.com/eriksonJAguilar/godiffpriv>

- Validar a usabilidade dos dados gerados pelos modelos de privacidade após serem modificados;
- Adicionar a implementação para outros formatos de dados, tais como texto e som. Além disso, integrar outros tipos de imagens médicas;
- Adicionar todas as regras da GDPR, bem como da HIPAA, para complementar as garantias de privacidade internacionais, haja vista que para esta proposta, foi utilizada apenas LGPD e alguns fragmentos da HIPAA.

REFERÊNCIAS

- AGUIAR, E. J. D.; FAIÇAL, B. S.; KRISHNAMACHARI, B.; UHEYAMA, J. A survey of blockchain-based strategies for healthcare. **ACM Computing Surveys**, Association for Computing Machinery, New York, NY, USA, v. 53, n. 2, mar. 2020. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/3376915>. Citado na página 41.
- AIELLO, M.; CAVALIERE, C.; D'ALBORE, A.; SALVATORE, M. The challenges of diagnostic imaging in the era of big data. **Journal of Clinical Medicine**, MDPI AG, v. 8, n. 3, p. 316, mar. 2019. Disponível em: <https://doi.org/10.3390/jcm8030316>. Citado na página 42.
- AL-ANI, A. K.; ANBAR, M.; MANICKAM, S.; AL-ANI, A. Dad-match; security technique to prevent denial of service attack on duplicate address detection process in ipv6 link-local network. **PLOS ONE**, Public Library of Science, v. 14, n. 4, p. 1–20, 04 2019. Disponível em: <https://doi.org/10.1371/journal.pone.0214518>. Citado na página 101.
- AL-DAHMAN, R. R.; SHI, Q.; LEE, G. M.; KIFAYAT, K. **Survey on revocation in ciphertext-policy attribute-based encryption**. [S.l.]: MDPI AG, 2019. Citado na página 45.
- AL-JAROODI, J.; MOHAMED, N. Blockchain in industries: A survey. **IEEE Access**, v. 7, p. 36500–36515, 2019. Citado na página 26.
- AL-OMAR, A.; BHUIYAN, M. Z. A.; KIYOMOTO, A. B.; SHINSAKU; RAHMAN, M. S. Privacy-friendly platform for healthcare data in cloud based on blockchain environment. **Future Generation Computer Systems**, Elsevier BV, v. 95, p. 511–521, jun. 2019. Disponível em: <https://doi.org/10.1016/j.future.2018.12.044>. Citado nas páginas 27, 64, 66, 67, 68, 69 e 103.
- ALHADHRAMI, Z.; ALGHFELI, S.; ALGHFELI, M.; ABEDLLA, J. A.; SHUAIB, K. Introducing blockchains for healthcare. In: **2017 International Conference on Electrical and Computing Technologies and Applications (ICECTA)**. IEEE, 2017. p. 1–4. Disponível em: <http://ieeexplore.ieee.org/document/8252043/>. Citado nas páginas 35 e 36.
- ALMADHOUN, N.; AYDAY, E.; ULUSOY Özgür. Inference attacks against differentially private query results from genomic datasets including dependent tuples. **Bioinformatics**, Oxford University Press (OUP), v. 36, n. Supplement_1, p. i136–i145, jul. 2020. Disponível em: <https://doi.org/10.1093/bioinformatics/btaa475>. Citado na página 103.
- ANDERSON, J. **Securing, standardizing, and simplifying electronic health record audit logs through permissioned blockchain technology**. Tese (Doutorado) — Dartmouth College, 2018. Disponível em: <https://www.cs.dartmouth.edu/~trdata/reports/TR2018-854.pdf>. Citado na página 26.
- ANDROULAKI, E.; BARGER, A.; BORTNIKOV, V.; CACHIN, C.; CHRISTIDIS, K.; CARO, A. D.; ENYEART, D.; FERRIS, C.; LAVENTMAN, G.; MANEVICH, Y.; MURALIDHARAN, S.; MURTHY, C.; NGUYEN, B.; SETHI, M.; SINGH, G.; SMITH, K.; SORNIOTTI, A.; STATHAKOPOULOU, C.; VUKOLIĆ, M.; COCCO, S. W.; YELICK, J. Hyperledger fabric: A distributed operating system for permissioned blockchains. In: **Proceedings of the Thirteenth**

EuroSys Conference. New York, NY, USA: ACM, 2018. (EuroSys '18), p. 30:1–30:15. ISBN 978-1-4503-5584-1. Disponível em: <http://doi.acm.org/10.1145/3190508.3190538>. Citado nas páginas 15, 37, 38, 39, 40, 56 e 75.

AXON, L.; GOLDSMITH, M.; CREESE, S. Privacy Requirements in Cybersecurity Applications of Blockchain. In: **Advances in Computers**. [S.l.]: Academic Press Inc., 2018. v. 111, p. 229–278. ISBN 9780128138526. Citado nas páginas 26, 27 e 73.

AZARIA, A.; EKBLAW, A.; VIEIRA, T.; LIPPMAN, A. MedRec: Using Blockchain for Medical Data Access and Permission Management. In: **2016 2nd International Conference on Open and Big Data (OBD)**. IEEE, 2016. p. 25–30. ISBN 978-1-5090-4054-4. Disponível em: <http://ieeexplore.ieee.org/document/7573685/>. Citado nas páginas 26, 54, 61, 66, 67, 68, 69 e 106.

BACH, L. M.; MISAKJEVIC; ZAGAR, M. Comparative Analysis of Blockchain Consensus Algorithms. In: **International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)**. Opatija, Croatia: [s.n.], 2018. p. 1545–1550. Disponível em: <https://ieeexplore.ieee.org/document/8400278>. Citado nas páginas 36 e 89.

BAGAI, R.; TANG, B.; KIM, E. Effectiveness of probabilistic attacks on anonymity of users communicating via multiple messages. **IEEE Systems Journal**, v. 7, n. 2, p. 199–210, 2013. Citado na página 48.

BANZI, R.; CANHAM, S.; KUCHINKE, W.; KRLEZA-JERIC, K.; DEMOTES-MAINARD, J.; OHMANN, C. Evaluation of repositories for sharing individual-participant data from clinical studies. **Trials**, Springer Science and Business Media LLC, v. 20, n. 1, mar. 2019. Disponível em: <https://doi.org/10.1186/s13063-019-3253-3>. Citado na página 25.

BARBOSA, P.; BRITO, A.; ALMEIDA, H. Privacy by Evidence: A Methodology to develop privacy-friendly software applications. **Information Sciences**, Elsevier Inc., v. 527, p. 294–310, jul 2020. ISSN 00200255. Citado nas páginas 48, 49, 58, 90, 91 e 103.

BARKER, W. C. **Guideline for identifying an information system as a national security system**. [S.l.], 2003. Acessado em: 19 de agosto de 2020. Disponível em: <https://doi.org/10.6028/nist.sp.800-59>. Citado na página 69.

BASSO, T.; MATSUNAGA, R.; MORAES, R.; ANTUNES, N. Challenges on anonymity, privacy, and big data. In: **2016 Seventh Latin-American Symposium on Dependable Computing (LADC)**. [S.l.: s.n.], 2016. p. 164–171. Citado na página 103.

BELLARE, M.; GOLDREICH, O. On defining proofs of knowledge. In: BRICKELL, E. F. (Ed.). **Advances in Cryptology — CRYPTO' 92**. Berlin, Heidelberg: Springer Berlin Heidelberg, 1993. p. 390–420. Citado na página 46.

BERTINO, E.; FOVINO, I. N.; PROVENZA, L. P. A framework for evaluating privacy preserving data mining algorithms. **Data Mining and Knowledge Discovery**, Springer Science and Business Media LLC, v. 11, n. 2, p. 121–154, ago. 2005. Disponível em: <https://doi.org/10.1007/s10618-005-0006-6>. Citado na página 59.

BLUM, M.; FELDMAN, P.; MICALI, S. Non-interactive zero-knowledge and its applications. In: **Proceedings of the Twentieth Annual ACM Symposium on Theory of Computing**. New York, NY, USA: Association for Computing Machinery, 1988. (STOC '88), p. 103–112. ISBN 0897912640. Disponível em: <https://doi.org/10.1145/62212.62222>. Citado na página 46.

BODKHE, U.; TANWAR, S.; PAREKH, K.; KHANPARA, P.; TYAGI, S.; KUMAR, N.; ALAZAB, M. Blockchain for Industry 4.0: A Comprehensive Review. **IEEE Access**, Institute of Electrical and Electronics Engineers (IEEE), v. 8, p. 79764–79800, apr 2020. ISSN 21693536. Citado nas páginas 31, 40, 41 e 44.

BOONSTRA, A.; VERSLUIS, A.; VOS, J. F. J. Implementing electronic health records in hospitals: a systematic literature review. **BMC Health Services Research**, Springer Science and Business Media LLC, v. 14, n. 1, set. 2014. Disponível em: <<https://doi.org/10.1186/1472-6963-14-370>>. Citado na página 25.

BRASIL, G. F. do. **Lei Geral de Privacidade de Dados Pessoais**. 2018. Acessado em: 10 de Junho de 2020. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm>. Citado nas páginas 42, 43, 44, 66 e 68.

BRITO, F. T.; MACHADO, J. C. Preservação de privacidade de dados: Fundamentos, técnicas e aplicações. **Anais da XXXVI Jornada de Atualização em Informática**, n. July, p. 91–130, 2017. Citado nas páginas 47 e 48.

BUTERIN, V. A next-generation smart contract and decentralized application platform. 2014. Acessado em: 17-06-2019 [White paper]. Disponível em: <<https://github.com/ethereum/wiki/wiki/White-Paper>>. Citado nas páginas 27, 33 e 38.

CACHIN, C.; VUKOLIC, M. Blockchain consensus protocols in the wild. **CoRR**, abs/1707.01873, 2017. Disponível em: <<http://arxiv.org/abs/1707.01873>>. Citado nas páginas 36, 37 e 38.

CAI, W.; WANG, Z.; ERNST, J. B.; HONG, Z.; FENG, C.; LEUNG, V. C. M. Decentralized applications: The blockchain-empowered software system. **IEEE Access**, v. 6, p. 53019–53033, 2018. ISSN 2169-3536. Disponível em: <<https://ieeexplore.ieee.org/document/8466786>>. Citado na página 38.

CASTRO, M.; LISKOV, B. Practical byzantine fault tolerance. In: **Proceedings of the Third Symposium on Operating Systems Design and Implementation**. Berkeley, CA, USA: USE-NIX Association, 1999. (OSDI '99), p. 173–186. ISBN 1-880446-39-1. Disponível em: <<http://dl.acm.org/citation.cfm?id=296806.296824>>. Citado nas páginas 56 e 66.

CAVOUKIAN, A. **Privacy by design: The 7 foundational principles**. 2009. Acessado em: 20 de agosto de 2020. Citado na página 58.

CHEN, B.-C.; KIFER, D.; LEFEVRE, K.; MACHANAVAJJHALA, A. Privacy-preserving data publishing. **Foundations and Trends in Databases**, v. 2, p. 1–167, 01 2009. Citado na página 47.

CHEN, Q. Toward realizing self-protecting healthcare information systems: Design and security challenges. In: **Advances in Computers**. Elsevier, 2019. p. 113–149. Disponível em: <<https://doi.org/10.1016/bs.adcom.2019.02.003>>. Citado na página 25.

CHUKWU, E.; GARG, L. A Systematic Review of Blockchain in Healthcare: Frameworks, Prototypes, and Implementations. **IEEE Access**, Institute of Electrical and Electronics Engineers (IEEE), v. 8, p. 21196–21214, jan 2020. Citado na página 86.

- CLARK, K.; VENDT, B.; SMITH, K.; FREYMAN, J.; KIRBY, J.; KOPPEL, P.; MOORE, S.; PHILLIPS, S.; MAFFITT, D.; PRINGLE, M. *et al.* The cancer imaging archive (tcia): maintaining and operating a public information repository. **Journal of digital imaging**, Springer, v. 26, n. 6, p. 1045–1057, 2013. Disponível em: <<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3824915/>>. Citado na página 84.
- CLIFTON, C.; ANANDAN, B. Challenges and opportunities for security with differential privacy. In: BAGCHI, A.; RAY, I. (Ed.). **Information Systems Security**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. p. 1–13. ISBN 978-3-642-45204-8. Citado na página 49.
- COMMISSION, E. **General Data Protection Regulation**. 2018. Acessado em: 10 de Junho de 2020. Disponível em: <<https://gdpr-info.eu/>>. Citado na página 42.
- CORTEN, P. **Implementation of Blockchain Powered Smart Contracts in Governmental Services**. Tese (Dissertação Mestrado) — Delft University of Technology, Delft, Países Baixos, 2018. Disponível em: <<http://resolver.tudelft.nl/uuid:87709465-b9a1-48da-9ba5-eba98bc263d7>>. Citado na página 54.
- COULOURIS, G.; DOLLIMORE, J.; KINDBERG, T.; BLAIR, G. **Sistemas Distribuídos - 5ed: Conceitos e Projeto**. [S.l.]: Bookman Editora, 2013. v. 5. ISBN 9788582600542. Citado na página 32.
- Dabbagh, M.; Sookhak, M.; Safa, N. S. The evolution of blockchain: A bibliometric study. **IEEE Access**, v. 7, p. 19212–19221, 2019. ISSN 2169-3536. Disponível em: <<https://ieeexplore.ieee.org/document/8628982>>. Citado na página 53.
- DANNEN, C. **Introducing Ethereum and Solidity: Foundations of Cryptocurrency and Blockchain Programming for Beginner**. Brooklyn, USA: [s.n.], 2017. Disponível em: <<https://link.springer.com/book/10.1007/978-1-4842-2535-6>>. Citado na página 38.
- DEWRI, R.; RAY, I.; RAY, I.; WHITLEY, D. On the optimal selection of k in the k-anonymity problem. In: **2008 IEEE 24th International Conference on Data Engineering**. [S.l.: s.n.], 2008. p. 1364–1366. Citado na página 80.
- DINH, T. T. A.; LIU, R.; ZHANG, M.; CHEN, G.; OOI, B. C.; WANG, J. Untangling blockchain: A data processing view of blockchain systems. **IEEE Transactions on Knowledge and Data Engineering**, v. 30, n. 7, p. 1366–1385, July 2018. ISSN 1041-4347. Disponível em: <<https://ieeexplore.ieee.org/document/8246573>>. Citado nas páginas 15, 33 e 34.
- DOH, O. D. N. **Managed Care Individual Provider Network Data State of New York**. 2016. Acessado em: 21 de agosto de 2020. Disponível em: <<https://health.data.ny.gov/Health/Managed-Care-Individual-Provider-Network-Data-Dece/ezbp-qg54>>. Citado na página 84.
- DUJAK, D.; SAJTER, D. Blockchain applications in supply chain. In: **SMART Supply Network**. Springer International Publishing, 2018. p. 21–46. Disponível em: <https://doi.org/10.1007/978-3-319-91668-2_2>. Citado na página 41.
- DWORK, C.; ROTH, A. The algorithmic foundations of differential privacy. **Foundations and Trends in Theoretical Computer Science**, Now Publishers Inc., Hanover, MA, USA, v. 9, p. 211–407, ago. 2014. ISSN 1551-305X. Disponível em: <<http://dx.doi.org/10.1561/04000000042>>. Citado nas páginas 49 e 79.

FABRIC, H. I. **Channels – Fabric Docs**. 2020. Acessado em: 03 de Fevereiro de 2021. Disponível em: <<https://hyperledger-fabric.readthedocs.io/en/release-2.2/channels.html>>. Citado na página 40.

_____. **Connection Options – Fabric Docs**. 2020. Acessado em: 05 de Fevereiro de 2021. Disponível em: <<https://hyperledger-fabric.readthedocs.io/en/release-2.2/developapps/connectionoptions.html>>. Citado nas páginas 92 e 93.

_____. **Fabric CA User’s Guide – Fabric Docs**. 2020. Acessado em: 04 de Fevereiro de 2021. Disponível em: <<https://hyperledger-fabric-ca.readthedocs.io/en/release-1.4/users-guide.html>>. Citado nas páginas 77 e 78.

_____. **Identity – Fabric Docs**. 2020. Acessado em: 20 de Agosto de 2020. Disponível em: <<https://hyperledger-fabric.readthedocs.io/pt/latest/identity/identity.html>>. Citado na página 40.

_____. **Private data – Fabric Docs**. 2020. Acessado em: 20 de Agosto de 2020. Disponível em: <<https://hyperledger-fabric.readthedocs.io/pt/latest/private-data/private-data.html>>. Citado na página 40.

_____. **Securing Communication With Transport Layer Security (TLS) – Fabric Docs**. 2020. Acessado em: 05 de Fevereiro de 2021. Disponível em: <https://hyperledger-fabric.readthedocs.io/en/release-2.2/enable_tls.html>. Citado na página 93.

FENG, Q.; DEBIAO, H.; ZEADALLY, S.; KHAN, M. K.; KUMAR, N. A survey on privacy protection in blockchain system. **Journal of Network and Computer Applications**, v. 126, p. 45–58, 2019. ISSN 1084-8045. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S1084804518303485>>. Citado nas páginas 26, 27, 41, 42, 44, 46, 47 e 67.

FLAUMENHAFT, Y.; BEN-ASSULI, O. Personal health records, global policy and regulation review. **Health Policy**, v. 122, n. 8, p. 815 – 826, 2018. ISSN 0168-8510. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0168851018301325>>. Citado nas páginas 26, 42 e 43.

GAN, C.; SAINI, A.; ZHU, Q.; XIANG, Y.; ZHANG, Z. Blockchain-based access control scheme with incentive mechanism for eHealth systems: patient as supervisor. **Multimedia Tools and Applications**, Springer Science and Business Media LLC, jul. 2020. Disponível em: <<https://doi.org/10.1007/s11042-020-09322-6>>. Citado nas páginas 27, 64, 66, 67, 68, 69 e 106.

GARFINKEL, S. L. **De-identification of personal information**. [S.l.], 2015. Acessado em: 19 de agosto de 2020. Disponível em: <<https://doi.org/10.6028/nist.ir.8053>>. Citado nas páginas 26, 27, 69 e 121.

GAŽI, P.; KIAYIAS, A.; ZINDROS, D. Proof-of-stake sidechains. In: **2019 IEEE Symposium on Security and Privacy (SP)**. [S.l.: s.n.], 2019. p. 139–156. Citado na página 37.

GE, Y.; NGUYEN, H.; ARCURY, T.; JOHNSON, A.; HWANG, W.; GAGE, H.; REYNOLDS, T.; CARR, J.; SANDBERG, J. Insight into the sharing of medical images. **Applied Clinical Informatics**, Georg Thieme Verlag KG, v. 03, n. 04, p. 475–487, 2012. Disponível em: <<https://doi.org/10.4338/aci-2012-06-ra-0022>>. Citado nas páginas 66 e 84.

GIL, A. C. **Como elaborar projetos de pesquisa**. [S.l.]: Atlas São Paulo, 2002. v. 4. Citado na página 51.

GOYAL, V.; PANDEY, O.; SAHAI, A.; WATERS, B. Attribute-based encryption for fine-grained access control of encrypted data. In: **Proceedings of the 13th ACM Conference on Computer and Communications Security**. New York, NY, USA: ACM, 2006. (CCS '06), p. 89–98. ISBN 1-59593-518-5. Disponível em: <http://doi.acm.org/10.1145/1180405.1180418>. Citado nas páginas 44 e 45.

GREVE, F.; SAMPAIO, L.; ABIJAUDE, J.; COUTINHO, A.; VALCY, Í.; QUEIROZ, S. Blockchain e a revolução do consenso sob demanda. **Minicursos do Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (Minicursos_SBRC)**, v. 36, 2018. Disponível em: <http://www.sbrc2018.ufscar.br/minicurso-5-mc-5/>. Citado nas páginas 33, 37 e 84.

GUO, R.; SHI, H.; ZHAO, Q.; ZHENG, D. Secure Attribute-Based Signature Scheme With Multiple Authorities for Blockchain in Electronic Health Records Systems. **IEEE Access**, v. 6, p. 11676–11686, 2018. ISSN 2169-3536. Disponível em: <http://ieeexplore.ieee.org/document/8279429/>. Citado nas páginas 27, 33, 63, 66, 67, 68 e 103.

HAFFKE, F. **Technical Analysis of Established Blockchain Systems**. Tese (Dissertação de mestrado) — Technical University of Munich, SW Engineering for Business Informatics, Munique, Alemanha, 2017. Disponível em: https://www.matthes.in.tum.de/file/jugj5lo6x15w/Sebis-Public-Website/Student-Theses-Guided-Research/Current-Bachelor-s-and-Master-s-Theses/Technical-Analysis-of-Established-Blockchain-Systems/HaffkeFlorian_KickOff_Final.pdf. Citado na página 33.

HOLUB, P.; KOZERA, L.; FLORINDI, F.; ENCKEVORT, E. van; SWERTZ, M.; REIHS, R.; WUTTE, A.; VALÍK, D.; MAYRHOFER, M. T. BBMRI-ERIC's contributions to research and knowledge exchange on COVID-19. **European Journal of Human Genetics**, Springer Science and Business Media LLC, v. 28, n. 6, p. 728–731, maio 2020. Disponível em: <https://doi.org/10.1038/s41431-020-0634-8>. Citado na página 25.

HOY, M. B. An Introduction to the Blockchain and Its Implications for Libraries and Medicine. **Medical Reference Services Quarterly**, v. 36, n. 3, p. 273–279, jul 2017. ISSN 0276-3869. Disponível em: <https://www.tandfonline.com/doi/full/10.1080/02763869.2017.1332261>. Citado na página 33.

HUGHES, D. R. Two concepts of privacy. **Computer Law & Security Review**, v. 31, n. 4, p. 527 – 537, 2015. ISSN 0267-3649. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0267364915000904>. Citado na página 42.

HöLBL, M.; KOMPARA, M.; KAMIŠALIĆ, A.; ZLATOLAS, L. N. A systematic review of the use of blockchain in healthcare. **Symmetry**, v. 10, n. 10, 2018. ISSN 2073-8994. Disponível em: <http://www.mdpi.com/2073-8994/10/10/470>. Citado na página 61.

ICHIKAWA, D.; KASHIYAMA, M.; UENO, T. Tamper-resistant mobile health using blockchain technology. **JMIR mHealth and uHealth**, JMIR Publications Inc., Toronto, Canada, v. 5, n. 7, 2017. Disponível em: <https://mhealth.jmir.org/2017/7/e111/>. Citado na página 56.

JAIN, R. **The art of computer systems performance analysis : techniques for experimental design, measurement, simulation, and modeling**. New York: Wiley, 1991. ISBN 0471503363. Citado nas páginas 96 e 97.

Jin, H.; Luo, Y.; Li, P.; Mathew, J. A review of secure and privacy-preserving medical data sharing. **IEEE Access**, v. 7, 2019. ISSN 2169-3536. Disponível em: <<https://ieeexplore.ieee.org/document/8713993>>. Citado nas páginas 26, 43, 84 e 85.

KHAN, S. I.; HOQUE, A. S. L. Secured technique for healthcare record linkage. In: **Proceedings of the 6th International Conference on Networking, Systems and Security**. New York, NY, USA: Association for Computing Machinery, 2019. (NSysS '19), p. 30–36. ISBN 9781450376990. Disponível em: <<https://doi.org/10.1145/3362966.3362972>>. Citado nas páginas 26 e 27.

KUO, T.-T.; OHNO-MACHADO, L. ModelChain: Decentralized Privacy-Preserving Healthcare Predictive Modeling Framework on Private Blockchain Networks. feb 2018. Disponível em: <<http://arxiv.org/abs/1802.01746>>. Citado na página 85.

KUSMIERZ, B.; STAUPPE, P.; GAL, A. **Extracting Tangle Properties in Continuous Time via Large-Scale Simulations**. [S.l.], 2018. Citado na página 31.

LAMPORT, L.; SHOSTAK, R.; PEASE, M. The byzantine generals problem. **ACM Trans. Program. Lang. Syst.**, ACM, New York, NY, USA, v. 4, n. 3, p. 382–401, jul. 1982. ISSN 0164-0925. Disponível em: <<http://doi.acm.org/10.1145/357172.357176>>. Citado na página 36.

LANGER, S. G.; TELLIS, W.; CARR, C.; DALY, M.; ERICKSON, B. J.; MENDELSON, D.; MOORE, S.; PERRY, J.; SHASTRI, K.; WARNOCK, M.; ZHU, W. The rsna image sharing network. **Journal of Digital Imaging**, v. 28, n. 1, p. 53–61, Feb 2015. Disponível em: <<https://doi.org/10.1007/s10278-014-9714-z>>. Citado na página 84.

LI, F.; LI, H.; NIU, B.; CHEN, J. Privacy Computing: Concept, Computing Framework, and Future Development Trends. **Engineering**, Elsevier Ltd, v. 5, n. 6, p. 1179–1192, dec 2019. ISSN 20958099. Citado nas páginas 41 e 42.

LI, Y.; YU, Y.; YANG, B.; MIN, G.; WU, H. Privacy preserving cloud data auditing with efficient key update. **Future Generation Computer Systems**, v. 78, p. 789 – 798, 2018. ISSN 0167-739X. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0167739X16302989>>. Citado na página 45.

LIANG, X.; ZHAO, J.; SHETTY, S.; LIU, J.; LI, D. Integrating blockchain for data sharing and collaboration in mobile healthcare applications. In: **2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)**. [s.n.], 2017. p. 1–5. ISSN 2166-9589. Disponível em: <<https://ieeexplore.ieee.org/document/8292361>>. Citado na página 56.

LIN, I.-C.; LIAO, T.-C. A Survey of Blockchain Security Issues and Challenges. **International Journal of Network Security**, v. 19, n. 55, p. 653–65901, 2017. Disponível em: <<http://ijns.femto.com.tw/>>. Citado nas páginas 33 e 41.

LINN, L. A.; KOO, M. B. **Blockchain for health data and its potential use in health it and health care related research**. 2016. Disponível em: <<https://www.healthit.gov/sites/default/files/11-74-ablockchainforhealthcare.pdf>>. Citado na página 26.

LIU, J.; LI, X.; YE, L.; ZHANG, H.; DU, X.; GUIZANI, M. BPDS: A Blockchain Based Privacy-Preserving Data Sharing for Electronic Medical Records. **2018 IEEE Global Communications Conference, GLOBECOM 2018 - Proceedings**, IEEE, p. 1–6, 2018. Citado na página 93.

- Lo, S. K.; Xu, X.; Chiam, Y. K.; Lu, Q. Evaluating suitability of applying blockchain. In: **2017 22nd International Conference on Engineering of Complex Computer Systems (ICECCS)**. [s.n.], 2017. p. 158–161. Disponível em: <https://ieeexplore.ieee.org/document/8292816>. Citado nas páginas 15, 54 e 55.
- LU, X.; AU, M. Chapter 11 - an introduction to various privacy models. In: AU, M. H.; CHOO, K.-K. R. (Ed.). **Mobile Security and Privacy**. Boston: Syngress, 2017. p. 235 – 245. ISBN 978-0-12-804629-6. Disponível em: <http://www.sciencedirect.com/science/article/pii/B9780128046296000110>. Citado nas páginas 48 e 49.
- MACHANAVAJJHALA, A.; KIFER, D.; GEHRKE, J.; VENKITASUBRAMANIAM, M. *l*-diversity: Privacy beyond *k*-anonymity. **ACM Trans. Knowl. Discov. Data**, Association for Computing Machinery, New York, NY, USA, v. 1, n. 1, p. 3–es, mar. 2007. ISSN 1556-4681. Disponível em: <https://doi.org/10.1145/1217299.1217302>. Citado nas páginas 48 e 80.
- MCGHIN, T.; CHOO, K.-K. R.; LIU, C. Z.; HE, D. Blockchain in healthcare applications: Research challenges and opportunities. **Journal of Network and Computer Applications**, v. 135, p. 62 – 75, 2019. ISSN 1084-8045. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1084804519300864>. Citado na página 53.
- MIAH, S. J.; GAMMACK, J.; HASAN, N. Methodologies for designing healthcare analytics solutions: A literature analysis. **Health Informatics Journal**, p. 1460458219895386, 2019. PMID: 31876227. Disponível em: <https://doi.org/10.1177/1460458219895386>. Citado na página 25.
- MINGXIAO, D.; XIAOFENG, M.; ZHE, Z.; XIANGWEI, W.; QIJUN, C. A review on consensus algorithm of blockchain. In: **EEE International Conference on Systems, Man, and Cybernetics, SMC 2017**. [s.n.], 2017. p. 2567–2572. ISBN 9781538616451. Disponível em: <https://ieeexplore.ieee.org/document/8123011>. Citado nas páginas 37 e 38.
- MUKHOPADHYAY, U.; SKJELLUM, A.; HAMBOLU, O.; OAKLEY, J.; YU, L.; BROOKS, R. A brief survey of cryptocurrency systems. In: **2016 14th Annual Conference on Privacy, Security and Trust (PST)**. [s.n.], 2016. p. 745–752. Disponível em: <https://ieeexplore.ieee.org/abstract/document/7906988>. Citado na página 33.
- NAKAMOTO, S. **Bitcoin: A peer-to-peer electronic cash system**. 2008. Acessado em: 01 de Junho de 2019 [White paper]. Disponível em: <https://bitcoin.org/bitcoin.pdf>. Citado nas páginas 32, 33, 37 e 53.
- NAOR, M.; YUNG, M. Universal one-way hash functions and their cryptographic applications. In: **Proceedings of the Twenty-first Annual ACM Symposium on Theory of Computing**. New York, NY, USA: ACM, 1989. (STOC '89), p. 33–43. ISBN 0-89791-307-8. Disponível em: <http://doi.acm.org/10.1145/73007.73011>. Citado na página 33.
- NASIR, Q.; QASSE, I. A.; TALIB, M. A.; NASSIF, A. B. Performance analysis of hyperledger fabric platforms. **Security and Communication Networks**, Hindawi Limited, v. 2018, p. 1–14, set. 2018. Disponível em: <https://doi.org/10.1155/2018/3976093>. Citado na página 86.
- NIELES, M.; DEMPSEY, K.; PILLITTERI, V. Y. **An introduction to information security**. [S.l.], 2017. Acessado em: 19 de agosto de 2020. Disponível em: <https://doi.org/10.6028/nist.sp.800-12r1>. Citado na página 69.

ONGARO, D.; OUSTERHOUT, J. In search of an understandable consensus algorithm. In: **2014 USENIX Annual Technical Conference (USENIX ATC'14)**. [s.n.], 2014. p. 305–319. Disponível em: <<https://www.usenix.org/system/files/conference/atc14/atc14-paper-ongaro.pdf>>. Citado na página 38.

OWASP, F. **Open Web Application Security Project (OWASP) – Top 10 Privacy Risks**. 2016. <<https://owasp.org/www-project-top-10-privacy-risks/>>. Acessado em: 20 de agosto de 2020. Citado na página 58.

_____. **Open Web Application Security Project (OWASP) – Top Ten Project**. 2017. <<https://owasp.org/www-project-top-ten/>>. Acessado em: 22 de agosto de 2020. Citado na página 91.

PATEL, V. A framework for secure and decentralized sharing of medical imaging data via blockchain consensus. **Health Informatics Journal**, p. 146045821876969, apr 2018. ISSN 1460-4582. Disponível em: <<http://journals.sagepub.com/doi/10.1177/1460458218769699>>. Citado nas páginas 54, 63, 66, 67, 68, 103 e 106.

POPOV, S. The Tangle. 2018. Acessado em: 01 de Junho de 2019 [White paper]. Disponível em: <https://iotatoken.com/IOTA_Whitepaper.pdf>. Citado nas páginas 31, 37, 38 e 53.

QUISQUATER, J.-J.; QUISQUATER, M.; QUISQUATER, M.; QUISQUATER, M.; GUILLOU, L.; GUILLOU, M. A.; GUILLOU, G.; GUILLOU, A.; GUILLOU, G.; GUILLOU, S. How to explain zero-knowledge protocols to your children. In: BRASSARD, G. (Ed.). **Advances in Cryptology — CRYPTO' 89 Proceedings**. New York, NY: Springer New York, 1990. p. 628–631. ISBN 978-0-387-34805-6. Disponível em: <<https://dl.acm.org/citation.cfm?id=118269>>. Citado na página 45.

RADANOVIĆ, I.; LIKIĆ, R. Opportunities for use of blockchain technology in medicine. **Applied Health Economics and Health Policy**, v. 16, n. 5, p. 583–590, Oct 2018. ISSN 1179-1896. Disponível em: <<https://doi.org/10.1007/s40258-018-0412-8>>. Citado nas páginas 53 e 61.

RAGHUPATHI, W.; RAGHUPATHI, V. Big data analytics in healthcare: promise and potential. **Health Information Science and Systems**, Springer Nature, v. 2, n. 1, fev. 2014. Disponível em: <<https://doi.org/10.1186/2047-2501-2-3>>. Citado na página 25.

RIFI, N.; RACHKIDI, E.; AGOULMINE, N.; TAHER, N. C. Towards using blockchain technology for eHealth data access management. In: **2017 Fourth International Conference on Advances in Biomedical Engineering (ICABME)**. IEEE, 2017. p. 1–4. ISBN 978-1-5386-1642-0. Disponível em: <<http://ieeexplore.ieee.org/document/8167555/>>. Citado na página 33.

RUFFING, T.; MORENO-SANCHEZ, P. ValueShuffle: Mixing confidential transactions for comprehensive transaction privacy in bitcoin. In: **Financial Cryptography and Data Security**. Springer International Publishing, 2017. p. 133–154. Disponível em: <https://doi.org/10.1007/978-3-319-70278-0_8>. Citado na página 47.

SADIKU, M. N.; MUSA, S. M. **Performance Analysis of Computer Networks**. [S.l.]: Springer Publishing Company, Incorporated, 2013. v. 1. ISBN 9783319016450. Citado na página 57.

SANKAR, L. S.; SINDHU, M.; SETHUMADHAVAN, M. Survey of Consensus Protocols on Blockchain Applications. In: **International Conference on Advanced Computing and**

Communication Systems. Coimbatore, India: [s.n.], 2017. ISBN 9781509045594. Citado na página 57.

SCOTT, D. W. Sturges' rule. **Wiley Interdisciplinary Reviews: Computational Statistics**, Wiley, v. 1, n. 3, p. 303–306, nov. 2009. Disponível em: <<https://doi.org/10.1002/wics.35>>. Citado na página 80.

SEH, A. H.; ZAROOUR, M.; ALENEZI, M.; SARKAR, A. K.; AGRAWAL, A.; KUMAR, R.; KHAN, R. A. Healthcare data breaches: Insights and implications. **Healthcare**, MDPI AG, v. 8, n. 2, p. 133, maio 2020. Disponível em: <<https://doi.org/10.3390/healthcare8020133>>. Citado na página 26.

SHAHNAZ, A.; QAMAR, U.; KHALID, A. Using blockchain for electronic health records. **IEEE Access**, Institute of Electrical and Electronics Engineers Inc., v. 7, p. 147782–147795, 2019. ISSN 21693536. Citado nas páginas 26 e 57.

SHEN, B.; GUO, J.; YANG, Y. MedChain: Efficient healthcare data sharing via blockchain. **Applied Sciences**, MDPI AG, v. 9, n. 6, p. 1207, mar. 2019. Disponível em: <<https://doi.org/10.3390/app9061207>>. Citado nas páginas 62, 66, 67, 68, 69 e 103.

SHI, S.; HE, D.; LI, L.; KUMAR, N.; KHAN, M. K.; CHOO, K.-K. R. Applications of blockchain in ensuring the security and privacy of electronic health record systems: A survey. **Computers & Security**, v. 97, p. 101966, 2020. ISSN 0167-4048. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S016740482030239X>>. Citado nas páginas 19, 61, 65 e 68.

SINGHAL, A.; WINOGRAD, T.; SCARFONE, K. A. **Guide to secure web services**. [S.l.], 2007. Disponível em: <<https://doi.org/10.6028/nist.sp.800-95>>. Citado na página 121.

SOLTANIAN, M. R. K.; AMIRI, I. S. Chapter 3 - problem solving, investigating ideas, and solutions. In: SOLTANIAN, M. R. K.; AMIRI, I. S. (Ed.). **Theoretical and Experimental Methods for Defending Against DDOS Attacks**. [S.l.]: Syngress, 2016. p. 33 – 45. ISBN 978-0-12-805391-1. Citado na página 100.

SOUSA, J.; BESSANI, A. From byzantine consensus to BFT state machine replication: A latency-optimal transformation. In: **2012 Ninth European Dependable Computing Conference**. IEEE, 2012. Disponível em: <<https://doi.org/10.1109/edcc.2012.32>>. Citado na página 63.

STALLINGS, W. **Criptografia e Segurança de Redes: Princípios e Práticas**. [S.l.]: Pearson, 2014. v. 6. ISBN 9788543014500. Citado nas páginas 32, 101 e 121.

STOUFFER, K.; ZIMMERMAN, T.; TANG, C.; LUBELL, J.; CICHONSKI, J.; MCCARTHY, J. **Cybersecurity framework manufacturing profile**. [S.l.], 2017. Disponível em: <<https://doi.org/10.6028/nist.ir.8183>>. Citado na página 121.

SU, S.; XIE, T.; Lü, S. A provably secure non-iterative hash function resisting birthday attack. **Theoretical Computer Science**, v. 654, p. 128 – 142, 2016. ISSN 0304-3975. Computing and Combinatorics. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0304397516001547>>. Citado na página 101.

SUN, H.; HUANG, Y. Distributed ledger technology and economic contract innovation. In: **Proceedings of the 3rd International Conference on Crowd Science and Engineering**. New York, NY, USA: ACM, 2018. (ICCSE'18), p. 15:1–15:7. ISBN 978-1-4503-6587-1. Disponível em: <<http://doi.acm.org/10.1145/3265689.3267929>>. Citado na página 31.

SWAN, M. **Blockchain: Blueprint for a new economy**. [S.l.]: "O'Reilly Media, Inc.", 2015. Citado nas páginas 26, 31, 33 e 35.

SWEENEY, L. k-anonymity: A model for protecting privacy. **International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems**, World Scientific Publishing Company, v. 10, n. 5, p. 557–570, oct 2002. ISSN 02184885. Citado nas páginas 47 e 103.

TANWAR, S.; PAREKH, K.; EVANS, R. Blockchain-based electronic healthcare record system for healthcare 4.0 applications. **Journal of Information Security and Applications**, Elsevier Ltd, v. 50, p. 102407, feb 2020. ISSN 22142126. Citado nas páginas 41, 83 e 86.

VIMERCATI, S. d. C. d.; FORESTI, S. Quasi-identifier. In: _____. **Encyclopedia of Cryptography and Security**. Boston, MA: Springer US, 2011. p. 1010–1011. ISBN 978-1-4419-5906-5. Disponível em: <https://doi.org/10.1007/978-1-4419-5906-5_763>. Citado na página 121.

WAGNER, I.; ECKHOFF, D. Technical privacy metrics: A systematic survey. **ACM Comput. Surv.**, Association for Computing Machinery, New York, NY, USA, v. 51, n. 3, jun. 2018. ISSN 0360-0300. Disponível em: <<https://doi.org/10.1145/3168389>>. Citado nas páginas 58, 59, 93, 94, 95 e 96.

WANG, J.; LI, M.; HE, Y.; LI, H.; XIAO, K.; WANG, C. A blockchain based privacy-preserving incentive mechanism in crowdsensing applications. **IEEE Access**, v. 6, p. 17545–17556, 2018. Citado na página 48.

WATERS, B. Ciphertext-policy attribute-based encryption: An expressive, efficient, and provably secure realization. In: CATALANO, D.; FAZIO, N.; GENNARO, R.; NICOLOSI, A. (Ed.). **Public Key Cryptography – PKC 2011**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. p. 53–70. ISBN 978-3-642-19379-8. Disponível em: <https://link.springer.com/chapter/10.1007/978-3-642-19379-8_4>. Citado na página 45.

WILL, M. A.; KO, R. K. Chapter 5 - a guide to homomorphic encryption. In: KO, R.; CHOO, K.-K. R. (Ed.). **The Cloud Security Ecosystem**. Boston: Syngress, 2015. p. 101 – 127. ISBN 978-0-12-801595-7. Disponível em: <<http://www.sciencedirect.com/science/article/pii/B9780128015957000057>>. Citado na página 47.

WU, H.; WANG, F. A Survey of Noninteractive Zero Knowledge Proof System and Its Applications. **The Scientific World Journal**, Hindawi Limited, v. 2014, 2014. Citado na página 46.

WUST, K.; GERVAIS, A. Do you need a blockchain? In: **2018 Crypto Valley Conference on Blockchain Technology (CVCBT)**. [s.n.], 2018. p. 45–54. Disponível em: <<https://ieeexplore.ieee.org/document/8525392>>. Citado nas páginas 15, 35, 36, 54 e 55.

XIA, Q.; SIFAH, E. B.; ASAMOAH, K. O.; GAO, J.; DU, X.; GUIZANI, M. MeDShare: Trust-Less Medical Data Sharing among Cloud Service Providers via Blockchain. **IEEE Access**, v. 5, p. 14757–14767, 2017. ISSN 21693536. Disponível em: <<https://ieeexplore.ieee.org/document/7990130>>. Citado nas páginas 62, 66, 67, 68 e 69.

XU, X.; WEBER, I.; STAPLES, M.; ZHU, L.; BOSCH, J.; BASS, L.; PAUTASSO, C.; RIMBA, P. A taxonomy of blockchain-based systems for architecture design. In: **2017 IEEE International Conference on Software Architecture (ICSA)**. [S.l.: s.n.], 2017. p. 243–252. Citado nas páginas 32 e 33.

YAGA, D.; MELL, P.; ROBY, N.; SCARFONE, K. **Blockchain technology overview**. [S.l.], 2018. Disponível em: <<https://doi.org/10.6028/nist.ir.8202>>. Citado nas páginas 32 e 35.

YUAN, Y.; WANG, F. Blockchain and cryptocurrencies: Model, techniques, and applications. **IEEE Transactions on Systems, Man, and Cybernetics: Systems**, v. 48, n. 9, p. 1421–1428, Sep. 2018. ISSN 2168-2216. Citado nas páginas 15, 76, 77, 78 e 106.

ZHANG, R.; XUE, R.; LIU, L. Security and privacy on blockchain. **ACM Computing Surveys**, ACM, New York, NY, USA, v. 52, n. 3, p. 51:1–51:34, jul. 2019. ISSN 0360-0300. Disponível em: <<http://doi.acm.org/10.1145/3316481>>. Citado nas páginas 41 e 47.

ZHENG, Z.; XIE, S.; DAI, H.-N.; CHEN, X.; WANG, H. Blockchain challenges and opportunities: A survey. **International Journal of Web and Grid Services**, Inderscience Publishers (IEL), v. 14, n. 4, p. 352–375, 2018. ISSN 1741-1106. Citado na página 32.

GLOSSÁRIO

Anonimização: é a técnica que remove identificadores de ligação presentes em bases de dados, esta é uma das técnicas de de-anonimização (GARFINKEL, 2015).

Ataque do aniversário: O ataque do aniversário, surge do paradoxo com mesmo nome, que infere qual a probabilidade de uma sala com n pessoas terem a mesma data de aniversário. Isto é análogo a um *hash*, ou seja, qual a probabilidade de encontrar um *hash* similar no conjunto de token (STALLINGS, 2014).

Ataques de ligação: é um tipo de ataque que tem o objetivo de re-identificar a identidade de indivíduos, a partir de um segundo conjunto de dados similar que contenha mais informações, assim cada registro é re-identificado (GARFINKEL, 2015).

Ataques de re-identificação: São ataques que tentam descobrir as identidades retiradas de dados públicos e discernir os padrões dos usuários utilizando de uma base de dados acessível pelo nó malicioso (GARFINKEL, 2015).

Framework: é uma abstração que reúne códigos comuns entre vários projetos de software provendo uma funcionalidade genérica. *framework* provê uma linguagem comum para gerenciar o sistema e padroniza a comunicação (STOUFFER *et al.*, 2017; SINGHAL; WINOGRAD; SCARFONE, 2007).

Pseudo-anonimização: também é uma técnica de de-anonimização que transforma identificadores e informações associadas a um indivíduo em pseudônimo (GARFINKEL, 2015).

Quasi-indentificador: são identificadores que combinados podem ser úteis para inferir a identidade de um indivíduo que está relacionado com o grupo formado pelos identificadores. Alguns desses identificadores são: raça, data de nascimento, sexo e código postal (VIMERCATI; FORESTI, 2011).

Técnicas de de-anonimização: São técnicas que removem a informações pessoais de coleções de dados, para evitar que sejam re-identificados por usuários malicioso quando disponibilizadas publicamente (GARFINKEL, 2015).

DIAGRAMA DE SEQUÊNCIA PARA O FLUXO DO *TOKEN* DE RASTREAMENTO

Este apêndice apresenta dois diagramas que representam o fluxo do processo de inserção do *token* e de auditoria dos *logs*. A inserção do *token* é importante para a rastreabilidade na blockchain, a Figura 17 ilustra como isso foi implementado. A Figura 18 descreve como o administrador da rede blockchain pode auditar *logs* e verificar quais foram os atores que tiveram contato com o ativo antes do vazamento.

Figura 17 – Diagrama de sequência que descreve o fluxo para o módulo de inserção do *token* no metadado da imagem DICOM.

Figura 18 – Diagrama de sequência que descreve o fluxo para auditar imagens vazadas com base no *token* contido nos metadados.