UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

ANDERSON FELLIPE DE VASCONCELOS VIDAL

APLICAÇÃO DE L-DIVERSIDADE NA ANONIMIZAÇÃO DE DADOS PÚBLICOS DA CAMPANHA DE VACINAÇÃO CONTRA COVID-19

CAMPINA GRANDE - PB

2021

ANDERSON FELLIPE DE VASCONCELOS VIDAL

APLICAÇÃO DE L-DIVERSIDADE NA ANONIMIZAÇÃO DE DADOS PÚBLICOS DA CAMPANHA DE VACINAÇÃO CONTRA COVID-19

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel ou Bacharela em Ciência da Computação.

Orientadora: Professora Dr. Carlos Eduardo Pires

CAMPINA GRANDE - PB

2021

ANDERSON FELLIPE DE VASCONCELOS VIDAL

APLICAÇÃO DE L-DIVERSIDADE NA ANONIMIZAÇÃO DE DADOS PÚBLICOS DA CAMPANHA DE VACINAÇÃO CONTRA COVID-19

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel ou Bacharela em Ciência da Computação.

BANCA EXAMINADORA:

Professor Dr. Carlos Eduardo Pires

Orientador – UASC/CEEI/UFCG

Professora Dr. Franklin de Souza Ramalho

Examinador – UASC/CEEI/UFCG

Professor Tiago Lima Massoni

Professor da Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em: 06 de ABRIL de 2022.

CAMPINA GRANDE - PB

RESUMO

A divulgação de dados é um processo que ocorre com o objetivo de trazer mais transparência e possibilitar análises de dados em geral. Visando garantir a privacidade dos dados, muitas divulgações são feitas anonimizando os registros (de banco de dados) a partir da remoção de informações que identifiquem os indivíduos envolvidos, como é o caso das divulgações dos dados públicos de vacinação contra a COVID-19. Porém, existem ataques que podem ser facilmente realizados em dados anonimizados apenas associando registros, através de atributos comuns com outras divulgações de dados com identificadores que não possuem informações sensíveis. Em razão disso, diversas técnicas de anonimização foram desenvolvidas, como por exemplo, a L-Diversidade. Este artigo tem como objetivo evidenciar o ganho de privacidade aplicando essa técnica sobre os dados de vacinação, em que foram realizados ataques de associação utilizando o perfil de beneficiários do PROUNI e dados públicos de agendamentos divulgados pela prefeitura municipal de Fortaleza. Como resultado, foi possível observar um aumento substancial na proteção de informações sensíveis..

Aplicação de L-Diversidade na anonimização de dados públicos de vacinação contra COVID-19

Anderson Fellipe de Vasconcelos Vidal
anderson.vidal@ccc.ufcg.edu.com
Universidade Federal de Campina Grande
Campina Grande, Paraíba

Carlos Eduardo Pires
cesp@csc.ufcg.edu.br
Universidade Federal de Campina Grande
Campina Grande, Paraíba

Thiago Nóbrega
thiagonobrega@gmail.com
Universidade Federal de Campina Grande
Campina Grande, Paraíba

RESUMO

A divulgação de dados é um processo que ocorre com o objetivo de trazer mais transparência e possibilitar análises de dados em geral. Visando garantir a privacidade dos dados, muitas divulgações são feitas anonimizando os registros (de banco de dados) a partir da remoção de informações que identifiquem os indivíduos envolvidos, como é o caso das divulgações dos dados públicos de vacinação contra a COVID-19. Porém, existem ataques que podem ser facilmente realizados em dados anonimizados apenas associando registros, através de atributos comuns com outras divulgações de dados com identificadores que não possuem informações sensíveis. Em razão disso, diversas técnicas de anonimização foram desenvolvidos como, por exemplo, a L-Diversidade. Este artigo tem como objetivo evidenciar o ganho de privacidade aplicando essa técnica sobre os dados de vacinação, em que foram realizados ataques de associação utilizando o perfil de beneficiários do PROUNI e dados públicos de agendamentos divulgados pela prefeitura municipal de Fortaleza. Como resultado, foi possível observar um aumento substancial na proteção de informações sensíveis.

1 INTRODUÇÃO

É comum que organizações, de diversas áreas de atuação, coletem e armazenem informações de pessoas como identificações, características físicas, dados financeiros, entre outros. Em específico, organizações públicas costumam divulgar informações a fim de tornar transparentes os processos públicos, como inscrições e resultados de concursos públicos e informações de vacinação.[1]

Com a crescente publicação de dados nos mais diversos meios de comunicação, também surgiu a preocupação com o conteúdo dos dados publicados e as consequências em divulga-los para as pessoas e a sociedade em geral. Dessa forma, visando garantir o direto a privacidade dos indivíduos, diversas estratégias de segurança foram desenvolvidas e implementadas no processo de divulgação dos dados públicos, como a ocultação, supressão e generalização de informação sensível[14].

Como exemplo a ser utilizado no decorrer deste trabalho, a publicação dos dados da campanha de vacinação contra COVID-19 [2] possui dados sobre a comorbidade de grupos de risco de indivíduos vacinados, sendo esta uma informação sensível e que não deve ser identificada por terceiros. Para garantir a privacidade dos indivíduos envolvidos, a divulgação destes dados utilizou a estratégia de ocultação, removendo dados de identificação dos indivíduos como o nome e o número de registro geral, ou seja, foram deixadas visíveis apenas informações genéricas e não sensíveis como, por exemplo, idade, gênero sexual e a sigla do estado de nascimento.

Quando a divulgação de dados não possui atributos sensíveis, como é o caso da divulgação do perfil de beneficiários do PROUNI [2], não existe a necessidade de ocultar atributos de identificação. Neste sentido, divulgações com essa característica são disponibilizados publicamente sem a ocultação de identificadores.

Porém, é possível utilizar da interseção entre os atributos dos registros de divulgações anonimizadas e não anonimizadas para associar uma identificação a uma informação sensível, desfazendo a anonimidade dos indivíduos e ferindo o direito a privacidade. Por exemplo, sabendo uma informação identificadora como nome de um indivíduo, podemos coletar seu registro em uma divulgação não anonimizada, possuindo as informações de idade, sexo, raça e local de nascimento, e buscar em uma divulgação anonimizada registros que possuam esse mesmo conjunto de informações, podendo encontrar apenas um único registro e então associar uma identificação a informação sensível.

Um estudo realizado por Sweeney [10], na Universidade Carnegie Mellon nos Estados Unidos, estimou ser possível identificar 87% da população de Massachusetts [7], incluindo o governador do estado. Para tal, bastou associar as informações de código postal, data de nascimento e gênero sexual dos eleitores do estado aos dados anônimos de visitas hospitalares de funcionários do estado publicados pela Comissão de Seguros de Grupo (GIC[3]), revelando suas condições médicas e outras informações privadas. A Figura 1 ilustra o diagrama de atributos comuns entre dos dados das visitas hospitalares e os registros de eleitores de Massachusetts.

Incursões como essas são chamados de ataques de associação ou Linkage Attack [8], em que é possível combinar conjuntos de dados através de atributos comuns para enriquecer informações acerca dos registros envolvidos, o que pode expor dados sensíveis de indivíduos ou organizações.

Com isso, diversas estratégias de anonimização foram desenvolvidas para combater este tipo de ataque, realizando modificações na forma como os atributos sensíveis e atributos não sensíveis são divulgados, utilizando como auxílio as diversas técnicas de

[2] Programa Universidade Para Todos (PROUNI)

[3] A Group Insurance Company (GIC) é responsável pela compra de seguro de saúde para os funcionários do estado americano de Massachusetts.

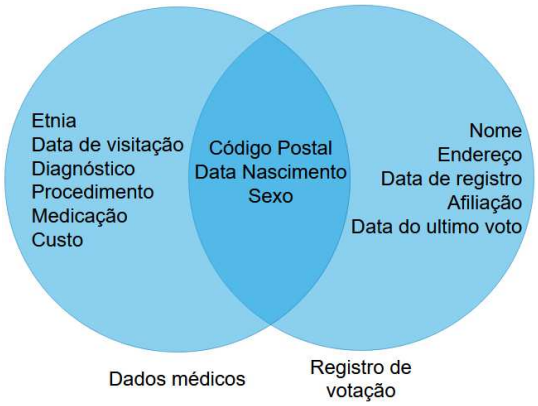

Figure 1: Interseção de atributos capaz de identificar indivíduos. Traduzida de [10].

supressão, agrupamento e inserção de ruído. Neste artigo serão descritos as técnicas de K-Anonimização [10] e o seu sucessor, a L-Diversidade [8] que tem como objetivo preservar a privacidade dos indivíduos agrupando registros características semelhantes e evitando re-identificações.

O objetivo deste trabalho é evidenciar as vulnerabilidades na divulgação de dados públicos sem o devido tratamento, realizando experimentos de ataques de associação aos dados da campanha de vacinação contra COVID-19 [2] utilizando dados divulgados pelo PROUNI [1] e de agendamentos de vacinação [3], a fim de identificar indivíduos.

Com o ataque a base de vacinação sem anonimização, foi possível observar que houve um número significativo de re-identificações, bem como ser possível agrupar cinco ou menos registros com a mesma característica, o que evidencia um alto risco de quebra de privacidade.

A partir disso, para minimizar e evitar a re-identificação de indivíduos, foi aplicada a técnica de anonimização L-Diversidade sobre os dados de vacinação [2]. Como resultado, ao repetir os ataques de associação, observou-se que não houve mais re-identificação, mostrando que a técnica é eficiente na preservação da privacidade dos indivíduos envolvidos.

Este artigo está estruturado da seguinte forma. Na Seção 2, é apresentada a fundamentação teórica. Na Seção 3, será discutida a metodologia aplicada para obtenção dos resultados. Na Seção 4, serão apresentadas as ferramentas que foram utilizadas para realização dos experimentos. Na seção 5, será apresentado os trabalhos relacionados que foram utilizados como base de conhecimento. E por fim, na seção 6, serão apresentadas as conclusões obtidas com os resultados apresentados, bem como a possibilidade de trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Esta seção fornece a fundamentação teórica para entendimento do assunto abordado neste trabalho. Primeiramente, é apresentado a técnica predecessora da L-Diversidade [8], chamada de K-Anonimização [10] e, em seguida são mostrados os princípios de privacidade a serem seguidos e o conceito de L-Diversidade.

2.1 K-Anonimização

Os ataques de associação fazem uso de atributos comuns entre os dados não sensíveis de uma fonte de dados externa e a base de dados anonimizada alvo do ataque. Estes atributos são chamados de quasi-identificadores [6], e é através deles que é possível associar a identificação de um indivíduo a um atributo sensíveis.

Com o objetivo de evitar os ataques de associação utilizando quasi-identificadores, foi desenvolvida uma técnica chamada K-Anonimização [10] que foca em generalizar ou suprimir de forma parcial estes atributos, criando grupos contendo k registros com os mesmos quasi-identificadores, o que dificulta os ataques de associação. Um agrupamento de dados será considerado k-anonimizado se para qualquer registro dentro do grupo, o mesmo seja indistinguível entre os k-1 outros registros do grupo que possuem o mesmo conjunto de valores para os atributos quasi-identificadores.

A Figura 2 corresponde a um quadro ilustrativo de registros de vacinação de indivíduos anônimos. Cada registro contém a idade, os cinco primeiros dígitos do código postal e o gênero como atributos não sensíveis. O grupo de vacinação representa o atributo sensível. Mesmo que esse conjunto de dados não esteja vinculado a um atributo identificador como, por exemplo, o nome do indivíduo, é possível identificar e descobrir o grupo de vacinação apenas conhecendo os valores dos atributos não sensíveis.

	Não sensível			Sensível
	Sexo	CEP	Idade	Grupo de vacinação
1	M	58441	29	DOENÇA RENAL CRÔNICA
2	F	59422	54	DIABETES MELLITUS
3	M	58461	28	DOENÇA CARDIOVASCULAR
4	M	58443	33	DIABETES MELLITUS
5	M	58460	31	DIABETES MELLITUS
6	F	59423	55	DOENÇA CARDIOVASCULAR
7	M	58469	38	DIABETES MELLITUS
8	F	59421	40	DOENÇA RENAL CRÔNICA
9	M	58442	26	DOENÇA CARDIOVASCULAR
10	F	59425	46	DIABETES MELLITUS
11	M	58467	29	DOENÇA RENAL CRÔNICA
12	M	58440	32	DIABETES MELLITUS

Figure 2: Dados ilustrativos de vacinação.

Aplicando o algoritmo de k-anonimização aos dados da Figura 2, devidamente configurado com k=4 (i.e., os grupos gerados terão quatro registros com os mesmos valores para os atributos quasi-identificadores), é possível observar na Figura 3 que foram gerados três grupos que possuem os mesmos quasi-identificadores. Para tal, suprimimos os dados de CEP e generalizamos os dados de idade.

Dessa forma, não é mais possível um simples ataque de associação vincular uma identificação a um único registro dessa tabela. Entretanto, podemos observar nos exemplos a seguir que ainda existem vulnerabilidades que podem ser exploradas mesmo em dados K-Anomizados.

Exemplo 1 - Ataque de homogeneidade: Artur e Pedro são

	Não sensível			Sensível
	Sexo	CEP	Idade	Grupo de vacinação
1	M	584**	> 30	DOENÇA RENAL CRÔNICA
2	M	584**	> 30	DOENÇA CARDIOVASCULAR
3	M	584**	> 30	DOENÇA RENAL CRÔNICA
4	M	584**	> 30	DOENÇA CARDIOVASCULAR
5	F	5942*	≤ 40	DOENÇA CARDIOVASCULAR
6	F	5942*	≤ 40	DIABETES MELLITUS
7	F	5942*	≤ 40	DIABETES MELLITUS
8	F	5942*	≤ 40	DOENÇA RENAL CRÔNICA
9	M	584**	3*	DIABETES MELLITUS
10	M	584**	3*	DIABETES MELLITUS
11	M	584**	3*	DIABETES MELLITUS
12	M	584**	3*	DIABETES MELLITUS

Figure 3: Dados ilustrativos de vacinação 4-anonimizado com base na Figura 2.

vizinhos próximos. Pedro sabe que Artur tem 32 anos e, como são vizinhos, conhece os cinco primeiros dígitos do seu CEP que é 58440. Pedro sabe que Arthur se vacinou recentemente e quer saber em qual grupo de vacinação ele foi registrado. Então Pedro acessa o microdado de vacinação 4-anonimizado da Figura 3 e encontra quatro registros (9, 10, 11 e 12) que satisfazem o conjunto de informações que possui. Pedro não consegue distinguir qual registro é de fato o registro de vacinação de Artur, porém, como todos os registros estão vinculados ao mesmo Grupo de vacinação, Pedro descobre que Artur está no grupo de vacinação de DIABETES MELLITUS e possui esta comorbidade.

Exemplo 2 - Ataque de conhecimento prévio: Pedro tem um amigo chamado Jorge, que conheceu pela Internet. Pedro sabe que ele se vacinou recentemente, que tem 29 anos e reside no CEP 58467. Acessando os dados da Figura 3, Pedro encontra quatro registros (1, 2, 3 e 4) que satisfazem o conjunto de informações que possui. Pedro agora sabe que Jorge pode ter uma doença renal ou cardiovascular. Contudo, Pedro lembra que o pai de Jorge é uma pessoa pública, e que possui um histórico de família de doenças renais. Logo, Pedro conclui que Jorge provavelmente possui uma doença renal.

Como é possível constatar, o algoritmo de k-anonimização não prevê a diversidade do atributo sensível tornando possível a existência de grupos homogêneos em relação à informação sensível, o que acaba expondo o indivíduo alvo do ataque. Além disso, a k-anonimização não protege os dados contra ataques baseados em conhecimentos externos como no caso do Exemplo 2. Em suma, o algoritmo de k-anonimização não garante a privacidade dos indivíduos. Para neutralizar estas vulnerabilidades, pode-se utilizar uma forma aprimorada da k-anonimização chamada L-Diversidade [8], a ser mostrada na Seção 2.3.

2.2 Princípio de privacidade

Para garantir a privacidade de um determinado conjunto de dados é importante levar em consideração alguns princípios que deveram ser satisfeitos com o algoritmo de L-Diversidade.

Existem duas formas de atacar a privacidade de um conjunto de dados anonimizado [8]. São elas:

Descoberta direta: Dado um conjunto de dados anonimizados, a descoberta direta acontece quando o atacante consegue identificar corretamente o valor sensível de um indivíduo com alta probabilidade. O ataque de homogeneidade do Exemplo 1 (Seção 2.1) é classificado como uma descoberta direta.

Descoberta indireta: Dado um conjunto de dados anonimizados, a descoberta indireta acontece quando o atacante consegue eliminar corretamente as possibilidades de valores sensíveis com alta probabilidade. O ataque de conhecimento prévio do Exemplo 2 (Seção 2.1) é classificado como uma descoberta indireta.

Com isso, é possível definir o princípio da uniformidade: a publicação de um conjunto de dados anonimizado deve oferecer pouca informação adicional ao conhecimento prévio de possíveis atacantes. Em outras palavras, não deve haver uma diferença expressiva entre o conhecimento anterior e posterior à publicação do conjunto de dados para o atacante. Ao satisfazer esse princípio para a publicação de dados, reduz-se as possibilidades de haver vazamentos de informações sensíveis de indivíduos.

2.3 L-Diversidade

A L-Diversidade [8] é uma técnica aprimorada da K-Anonimização para anonimizar conjuntos de dados públicos. Seu algoritmo introduz aos dados mais entropia/diversidade, fazendo com que os agrupamentos de dados, cada um representado pelo mesmo conjunto de quasi-identificadores, apresentem uma homogeneidade mínima.

Um conjunto de dados satisfaz L-Diversidade se, para cada agrupamento de registros que compartilham uma combinação de quasi-identificadores, existem pelo menos L valores “bem representados” para cada atributo sensível. Usando os dados da Figura 2, e aplicando o algoritmo de L-Diversidade para L=3, tem-se o resultado ilustrado na Figura 4.

É possível entender um agrupamento de registros como "bem representado" de três formas diferentes para o contexto de L-Diversidade, a citar:

Distinção L-Diversidade: deve haver pelo menos L valores distintos dos atributos sensíveis dentro de um agrupamento de dados com os mesmos quasi-identificadores. Comparando os dados mostrados nas Figura 3 e 4, percebe-se que os grupos da Figura 4 possuem três ou mais valores sensíveis distintos, diferentemente daqueles mostrados na Figura 3 onde há grupos nos quais os valores sensíveis são mais homogêneos.

Entropia L-Diversidade: a entropia de um agrupamento de dados define o quão diverso são os valores de um atributo sensível dentro do grupo. Formalmente, é possível definir a entropia de um

	Não sensível			Sensível
	Sexo	CEP	Idade	Grupo de vacinação
1	M	5844*	> 40	DOENÇA RENAL CRÔNICA
2	M	5844*	> 40	DOENÇA CARDIOVASCULAR
3	M	5844*	> 40	DIABETES MELLITUS
4	M	5844*	> 40	DIABETES MELLITUS
5	F	5942*	≤ 40	DIABETES MELLITUS
6	F	5942*	≤ 40	DOENÇA CARDIOVASCULAR
7	F	5942*	≤ 40	DIABETES MELLITUS
8	F	5942*	≤ 40	DOENÇA RENAL CRÔNICA
9	M	5846*	> 40	DIABETES MELLITUS
10	M	5846*	> 40	DIABETES MELLITUS
11	M	5846*	> 40	DOENÇA RENAL CRÔNICA
12	M	5846*	> 40	DOENÇA CARDIOVASCULAR

Figure 4: Dados ilustrativos de vacinação 3-diverso com base na Figura 2.

agrupamento de dados pela seguinte equação:

$$H(G) = -\sum_{c \in C} p(G,s) log(p(G,s)) \quad (1)$$

onde p(G,s) é a probabilidade de um registro ter o valor de atributo sensível s no grupo G. Pode-se dizer que um conjunto de dados possui entropia L-diversidade se, para cada grupo G, a Entropia $H(G) \geq log(L)$.

A consequência desta condição é que, para cada agrupamento, vão existir L valores distintos do atributo sensível. Com base nesta definição, e utilizando a equação de entropia, podemos dizer que os dados da Figura 4 é 2.8-diverso. Observe que a entropia L-diversidade captura a noção de grupos bem representados devido ao fato de que a entropia aumenta à medida em que as frequências de cada valor sensível ficam mais uniformes.

Rercusivo (c,L)-Diversidade: diz-se que um conjunto de dados é (c,L)-Diverso se, para cada agrupamento, removendo os registros que possuem um valor sensível s, ainda se mantém a diversidade dos valores sensíveis do agrupamento ((L-1)-Diverso). É uma maneira de garantir que os valores mais recorrentes não apareçam com muita frequência, e que os valores menos recorrentes não apareçam muito raramente nos grupos gerados pelo algoritmo.

3 METODOLOGIA

Para realização desta pesquisa, foram utilizadas duas publicações de dados públicos da campanha de vacinação contra a COVID-19 [2] referentes aos estados da Paraíba e do Ceará. As publicações que serão alvo dos testes de ataque e anonimização utilizando o algoritmo de L-Diversidade foram disponibilizadas pelo Ministério da Saúde (MEC).

Para associação com os dados de vacinação, foram utilizados como dados externos os dados públicos de estudantes bolsistas beneficiários do PROUNI [4] [1], disponibilizados pelo MEC, e dados de agendamento de vacinação da primeira dose do estado do Ceará [3], referentes aos meses de Julho a Setembro de 2021, disponibilizados pela Secretaria Municipal de Saúde de Fortaleza.

O ataque de associação foi desenvolvido através de script [11][12], no qual foram realizadas consultas as bases dados externas e à base de dados alvo, com o objetivo de associar os atributos quasi-identificadores de ambas as bases e realizar identificações, como também reconhecer o menor grupo com os mesmos valores quasi-identificadores.

A seguir, são detalhadas as características dos dados utilizados, as anonimizações realizadas e os algoritmos de ataque utilizados.

3.1 Dados Alvo: Campanha de Vacinação

Para demonstração das vulnerabilidades de privacidade dos dados e do benefício da anonimização de dados públicos utilizando L-Diversidade, foram usados os dados da Campanha de Vacinação. Tais dados estão divididos em duas bases de dados, referentes aos estados da Paraíba e do Ceará, com os registros de vacinação coletados do dia 17 de Janeiro de 2021 até o dia 15 de Janeiro de 2022. Ambas as bases de dados possuem os mesmos atributos, mostrados na Figura 5.

Para que a base de dados de vacinação pudesse ser utilizada, foi necessário remover registros com dados vazios ou nulos. Além disso, os dados foram reduzidos para apenas os registros referentes à aplicação da primeira dose, a fim de evitar que um mesmo indivíduo tenha mais de um registro na na base de dados, já que a vacinação ocorreu mais de uma vez para uma mesma pessoa. Foi escolhida a primeira dose devido à quantidade de registros (3.128.568 registros) ser maior que as da 2ª(2.524.555 registros).

3.2 Dados Externos: PROUNI e Agendamentos

Para realização da correlação com os dados de vacinação, foram utilizados como dados externos os registros de bolsas do PROUNI, cujos atributos podem ser visualizados na Figura 6, e os registros de agendamento da primeira dose da vacina no estado do Ceará.

A base de dados de agendamentos foi construída a partir da leitura de três arquivos em formato PDF utilizando um script [13] para gerar um único arquivo em CSV. Os arquivos PDF utilizados possuíam registros de agendamentos da primeira dose, referentes aos meses de julho, agosto e setembro de 2021. Os atributos da base dados gerada podem ser visualizados na Figura 7.

Para fins de análise, foram excluídos registros de agendamentos que possuíam inconsistências como valores nulos e vazios. Já para os dados do PROUNI não houve qualquer alteração.

3.3 Associação dos Dados de Vacinação com Dados Externos

Para realizar a associação dos dados externos com os dados de vacinação foi necessário identificar os atributos quasi-identificadores em ambas as bases de dados, ou seja, aqueles que possuem a mesma natureza e objetivo. Neste sentido, foram identificados os atributos

[4] Programa Universidade para Todos

Atributos da tabela da campanha de vacinação	
Atributo	Tipo
document_id	texto
paciente_id	texto
paciente_idade	decimal
paciente_dataNascimento	texto
paciente_enumSexoBiologico	decimal
paciente_racaCor_codigo	texto
paciente_racaCor_valor	texto
paciente_endereco_coIbgeMunicipio	texto
paciente_endereco_coPais	texto
paciente_endereco_nmMunicipio	texto
paciente_endereco_nmPais	texto
paciente_endereco_uf	texto
paciente_endereco_cep	texto
paciente_nacionalidade_enumNacionalidade	texto
estabelecimento_valor	inteiro
estabelecimento_razaoSocial	texto
estalecimento_noFantasia	texto
estabelecimento_municipio_codigo	inteiro
estabelecimento_municipio_nome	texto
estabelecimento_uf	texto
vacina_grupoAtendimento_codigo	inteiro
vacina_grupoAtendimento_nome	texto
vacina_categoria_codigo	decimal
vacina_categoria_nome	texto
vacina_lote	texto
vacina_fabricante_nome	texto
vacina_fabricante_referencia	texto
vacina_dataAplicacao	texto
vacina_descricao_dose	texto
vacina_codigo	inteiro
vacina_nome	texto
sistema_origem	texto

Figure 5: Atributos da tabela da campanha de vacinação.

referentes ao gênero, raça, data de nascimento, nome do município e sigla do estado nos dados do PROUNI, para fazer associação com os atributos da base de dados de vacinação da Paraíba. A Figura 8 mostra a correspondência destes atributos com os atributos dos dados de vacinação.

Já para a associação dos dados de vacinação com os dados de agendamento, foram identificados os atributos referentes à dose, data de nascimento e data de aplicação da vacina. No entanto, o

Atributos da tabela PROUNI	
Atributo	Tipo
ANO_CONCESSAO_BOLSA	texto
CODIGO_EMEC_IES_BOLSA	texto
NOME_IES_BOLSA	texto
MUNICIPIO	texto
CAMPUS	texto
TIPO_BOLSA	texto
MODALIDADE_ENSINO_BOLSA	texto
NOME_CURSO_BOLSA	texto
NOME_TURNO_CURSO_BOLSA	texto
CPF_BENEFICIARIO	texto
SEXO_BENEFICIARIO	texto
RACA_BENEFICIARIO	texto
DATA_NASCIMENTO	texto
BENEFICIARIO_DEFICIENTE_FISICO	texto
REGIAO_BENEFICIARIO	texto
UF_BENEFICIARIO	texto
MUNICIPIO_BENEFICIARIO	texto

Figure 6: Atributos da tabela do PROUNI.

Atributos da tabela de agendamentos	
Atributo	Tipo
nome	texto
data_nascimento	texto
localvacinacao	texto
data	texto
hora	texto
dose	texto

Figure 7: Atributos da tabela de agendamentos de vacinação do Ceará.

atributo dose foi desconsiderado devido à filtragem realizada nos dados de vacinação para selecionar apenas os registros referentes à primeira dose. Dessa forma, a correlação se limitou apenas a dois atributos, como mostra a figura 9.

3.4 Estratégia de Ataque e Proteção dos dados

Para gerar os resultados deste artigo e mostrar o ganho de privacidade de dados com o uso do L-Diversidade nos dados públicos de vacinação, foram realizados ataques de associação. Primeiramente, os ataques ocorreram nos dados de vacinação sem anonimização a fim de demonstrar a vulnerabilidade mostrando a quantidade de registros processados, identificados e não identificados, bem como a taxa de re-identificação.

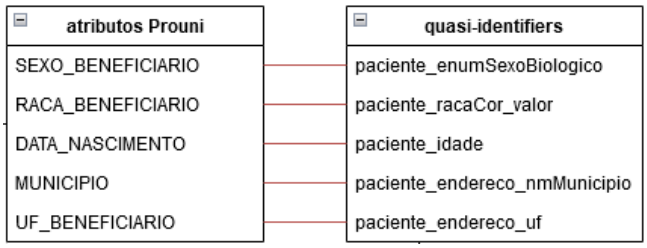

Figure 8: Associação entre os atributos da tabela do PROUNI e os quasi-identificadores da tabela de vacinação.

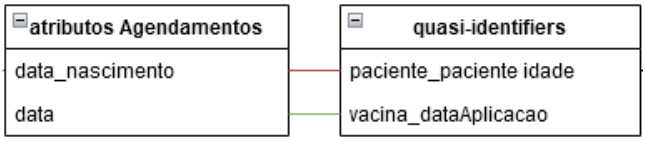

Figure 9: Associação entre os atributos da tabela de Agendamentos e os quasi-identificadores da tabela de vacinação.

Em seguida, utilizando a aplicação ARX [9], o atributo sensível foi identificado (vacina_grupoAtendimento_nome) e anonimizado utilizando o algoritmo de L-Diversidade para L com os valores 5, 10, 15 e 20. Para isso, foi necessário aplicar aos atributos quasi-identificadores, níveis de generalização para possibilitar a geração dos agrupamentos de dados. Os níveis de generalização para os atributos de data de nascimento podem ser verificados na Figura 10. O atributo paciente_idade foi suprimido completamente pois, como a data de nascimento está sendo generalizada, o atributo de idade poderia ser usado para contornar esta generalização.

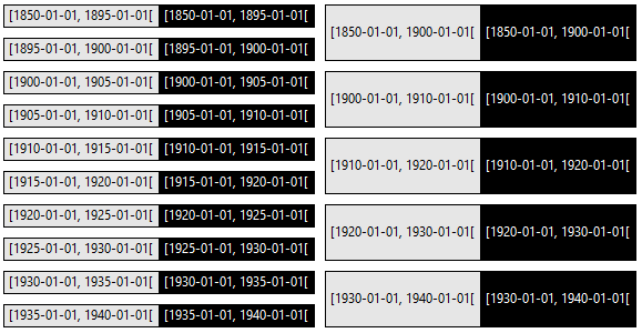

Figure 10: Representação parcial dos níveis de generalização para o atributo paciente_dataNascimento da tabela de vacinação.

O uso do atributo data de nascimento como quasi-identificador, ao invés da idade, se deu pela especificidade semântica do mesmo. Enquanto a idade considera apenas o ano de nascimento, a data de nascimento traz informações mais específicas que são dia e o mês, o que facilita possíveis re-identificações.

A partir dos dados anonimizadas geradas, foram realizados os ataques de associação e coletados os resultados. Para facilitar o processamento dos dados, foi necessário realizar a indexação dos dados de vacinação utilizando o conjunto de quasi-identificadores como índice dos registros, o que proporcionou à tarefa de ataque um melhor desempenho.

4 EXPERIMENTOS E RESULTADOS

Esta seção descreve as bases de dados e ferramentas utilizadas nos experimentos. Em seguida, são apresentados os experimentos e resultados obtidos, bem como uma discussão sobre os mesmos. Os experimentos consistem em realizar ataques de associação sobre os dados de vacinação com e sem anonimização, utilizando os dados do PROUNI e de Agendamentos. No primeiro caso, objetivo é evidenciar a vulnerabilidade de segurança da informação, enquanto que no segundo caso a ideia é inferir sobre a preservação da privacidade dos dados anonimizados utilizando a técnica da L-Diversidade.

4.1 Bases de dados

Para realização dos experimentos de ataque de associação, foram utilizadas como alvo as bases de dados de vacinação do estado da Paraíba que possui cerca de 3GB e pouco mais de 6 milhões de registros de indivíduos, e do Ceará que possui cerca de 6GB e pouco menos de 13 milhões de registros. Ambas as bases foram reduzidas a apenas registros referentes à aplicação da primeira de vacinação.

Além disso, foi utilizada a base de dados do PROUNI [1], que possui cerca de 400KB e aproximadamente 2.800 registros, para realização dos ataques aos registros de vacinação da Paraíba. Também foram utilizados registros de agendamentos de vacinação no município de Fortaleza, construída a partir da leitura de três arquivos em formato PDF, referentes a aplicação da primeira dose nos meses de Julho, Agosto e Setembro de 2021. Para tal, foi utilizado um script [13] desenvolvido na linguagem Python 3.9.0 juntamente com a biblioteca Tika 1.28.1, o que gerou um único CSV.

4.2 Script de Ataque

Para realização dos ataques de associação, foram utilizados os scripts [11] e [12]. Os scripts foram executados utilizando a aplicação Jupyter Notebook 6.4.5 com a biblioteca Pandas 1.4.1. Além disso, foi utilizada a aplicação gratuita ARX 3.9.0 [9], desenvolvida em Java, para realizar as anonimizações dos dados a partir do algoritmo de L-Diversidade, bem como supressão e generalização dos dados quasi-identificadores

4.3 Resultados

A seguir, são apresentados os resultados obtidos mediante a realização de ataques de associação a base de dados alvo (vacinação) com e sem anonimização dos dados.

4.3.1 Ataque de Associação à aos dados não Anonimizados. Na realização do ataque de associação aos dados da campanha de vacinação da Paraíba utilizando os dados dos beneficiários do PROUNI, foi coletado a quantidade de registros identificados, vinculando o atributo 'nome' dos registros do PROUNI, ao atributo sensível vacina_grupoVacinação_nome do referente registro de vacinação, para re-identificação. Além disso, também foram coletados o número de registros que possuíam os mesmos quasi-identificadores em grupos de até cinco registros para identificar a possibilidade de ataques

de homogeneidade. Os dados obtidos podem ser observados nas Figuras 11 e 12.

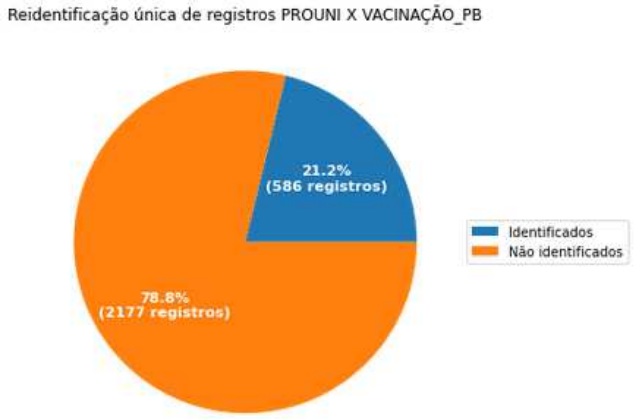

Figure 11: Reidentificação única dos grupos de vacinação da Paraíba para os registros do PROUNI.

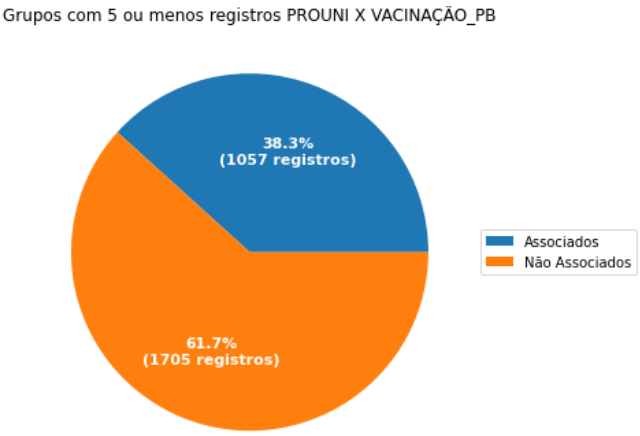

Figure 12: Grupos de cinco ou menos registros com os mesmos quasi-identificadores nos dados de vacinação da Paraíba.

É possível observar que a re-identificação dos indivíduos registrados na base de dados de vacinação é bastante expressiva sendo possível mapear 586 indivíduos aos seus respectivos grupos de vacinação. Isso indica que o dado sensível que pode revelar a comorbidade dos indivíduos não está anonimizado. Ao analisar o número de grupos com cinco ou menos registros com os mesmos quasi-identificadores, pode-se observar que, além das identificações, ainda existe um grande risco de re-identificação utilizando ataques de homogeneidade.

Da forma similar, para a base de dados da campanha de vacinação do Ceará, ao utilizar dados de agendamentos de vacinação (primeira dose) e realizar a re-identificação e contagem de grupos com cinco ou menos registros contendo os mesmos quasi-identificadores, foram obtidos os resultados mostrados nas Figuras 13 e 14.

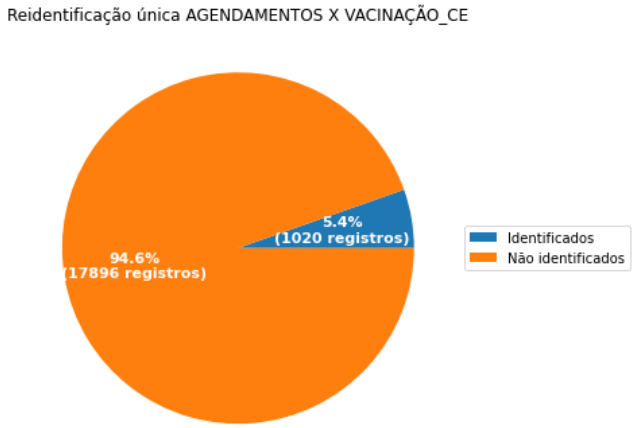

Figure 13: Reidentificação única dos grupos de vacinação do Ceará para os registros do agendamento.

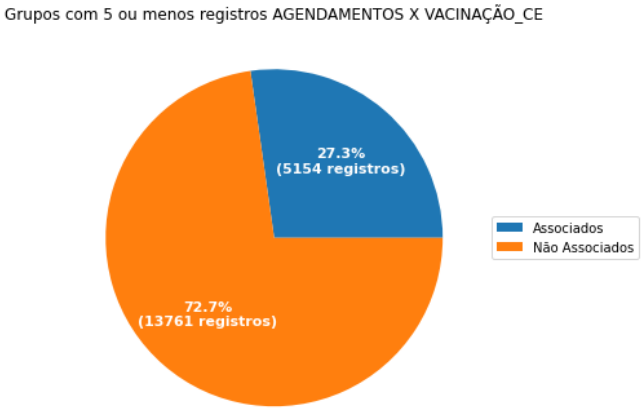

Figure 14: Grupos de cinco ou menos registros com os mesmos quasi-identificadores nos dados de vacinação do Ceará.

A re-identificação retornou apenas 5,4% de indivíduos identificados, porém representa ainda um número expressivo de 1020 pessoas cujos grupos de vacinação foram revelados. Quando analisado o número de grupos de até cinco registros com os mesmos quasi-identificadores, podemos observar que o risco de ataques de homogeneidade atinge em mais de 4000 pessoas.

4.4 Ataque de Associação aos dados Anonimizados

Com a anonimização dos dados de campanha de vacinação da Paraíba e do Ceará, utilizando o algoritmo de L-Diversidade, espera-se que as vulnerabilidades expostas no experimento anterior sejam sanadas ou ao menos minimizadas para ataques de associação.

A associação única (re-identificação) entre os registros das bases de dados externas e a base de dados alvore-identificação direta se tornou impossível devido à anonimização aplicada. Assim, foi coletado o menor grupo de registros anonimizados com os mesmos quasi-identificadores da campanha de vacinação. O gráfico da Figura 15 mostra o menor número de registros necessário para estimar

um possível risco de ataque de homogeneidade à base de dados de vacinação L-Diversificada da Paraíba em relação ao valor da constante L, os quais variam de cinco em cinco, de 5 até 20.

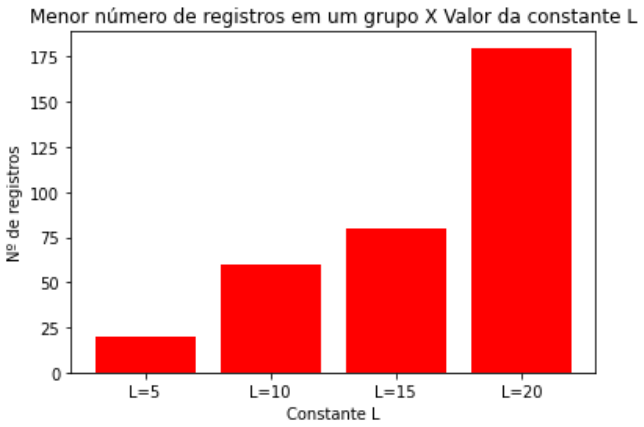

Figure 15: Relação entre o número de registros mínimos para identificação e o valor da constante L do algoritmo de L-Diversidade.

Pode-se observar que o acréscimo do valor da constante L leva ao aumento do número mínimo de registros com os mesmos quasi-identificadores, o que implica que o ataque de homogeneidade por associação se torna cada vez mais difícil à medida em que é aumentada a diversidade dos atributos sensíveis nos grupos.

Os ataques à base de dados anonimizada da campanha de vacinação do Ceará também não tiveram nenhum efeito expressivo, ou seja, não houve qualquer re-identificação e o ataque à base anonimizada 5-Diversa mostrou um número de registros contendo os mesmos quasi-identificados acima de 3.500 registros, impossibilitando a re-identificação através ataques de homogeneidade por associação.

Deve-se considerar também que a anonimização e o ataque levaram em consideração apenas dois quasi-identificadores: data de nascimento e data de vacinação, enquanto que o ataque utilizando os dados do PROUNI empregou cinco quasi-identificadores. Isto significa que o ataque poderia ser mais eficiente caso houvesse uma maior número de quasi-identificadores a serem considerados.

5 TRABALHOS RELACIONADOS

Para condução desta pesquisa, foram analisados diversos trabalhos acadêmicos sobre algoritmos de anonimização e conceitos de privacidade de dados. A seguir, descrevemos os principais artigos que nortearam a realização deste trabalho.

Os artigos [10] e [7] descrevem de forma clara e objetiva as vulnerabilidades nas publicações de dados realizadas pelos governos dos estados norte-americanos. O artigo [10] propõe a solução K-Anonimity e introduz o conceito de quasi-identificador, em que propõe realizar generalizações e supressões nos atributos quasi-identificadores a fim de criar grupos de k registros contendo o mesmo conjunto de valores para estes atributos.

Posteriormente, foi desenvolvida a técnica de anonimização derivada da K-Anonimity denominada de L-Diversity [8], em que foram descritos os detalhes teóricos acerca da privacidade dos indivíduos em publicações de dados, detalhando os conceitos matemáticos que cercam o algoritmo da L-Diversidade.

6 CONCLUSÃO E TRABALHOS FUTUROS

A partir das análises dos dados gerados pelos ataques de associação aos dados públicos de vacinação, foi observado que é possível extrair informações privadas dos indivíduos a partir de atributos aparentemente irrelevantes como gênero sexual, raça, nome do município e estado onde o indivíduo reside além de informações como data de nascimento e idade. Com isso, pode-se constatar que a divulgação dos dados públicos da campanha de vacinação contra COVID-19, mesmo sem atributos identificadores como nome do indivíduo e CPF, não garante a anonimidade das informações privadas dos cidadãos presentes nos dados da campanha.

A falta de proteção dos dados de vacinação torna-o vulnerável à realização de ataques simples, como os de associação, indo de encontro à Lei Geral de Proteção de Dados Pessoais[5] [5] que regulamenta os requisitos necessários para garantir a preservação da privacidade na publicação de dados públicos que envolvem informações sensíveis de indivíduos.

Como vimos no decorrer deste trabalho, para anonimizar os dados da campanha de vacinação, foi necessário identificar os atributos quasi-identificadores. Além disso, foi preciso aplicar a cada um deles generalizações e/ou supressões, para que fosse possível criar os agrupamentos de registros com os mesmos valores de quasi-identificadores, o que dificulta a associação com dados externos. Porém, generalizar ou suprimir os dados também prejudica análises posteriores para identificar informações importantes.

Por exemplo, se suprimíssemos os dados de gênero sexual, não seria possível identificar se diferenças entre a quantidade de vacinas aplicadas a pessoas do gênero masculino ou feminino, e então extrair conclusões acerca da aplicação da vacina na sociedade.

A anonimização utilizando o algoritmo de L-Diversidade não torna os dados totalmente imunes a outros tipos de ataque como, por exemplo, ataques de assimetria e semelhança [4], os quais conseguem extrair informações sensíveis de indivíduos mesmo em dados anonimizados. Isso ocorre porque o algoritmo L-Diversidade não prevê a relação semântica entre o conjunto de valores de um atributo sensível. Assim, um trabalho futuro consiste em propor técnicas para tratar tais vulnerabilidades do algoritmo L-Diversidade.

7 AGRADECIMENTOS

Agradeço primeiramente a Deus por sempre me guiar e renovar minhas forças em cada etapa de minha vida. Gratidão à minha família pelo apoio e por acreditarem que este trabalho fosse possível. Ao meu orientador Carlos Eduardo e o aluno de Doutorado Thiago Nóbrega, pela paciência e dedicação de seu tempo no acompanhamento deste trabalho. Aos meus colegas de trabalho e de meio acadêmico que me deram todo o apoio para que este trabalho pudesse ser feito. E aos demais professores do curso que fazem parte da minha formação como profissional e como integrante da sociedade.

[5] LGPD - Lei geral de proteção de dados - Lei n. 13.709, de 14 de agosto de 2018

REFERENCES

[1] 2020. Dados das bolsas e perfil dos beneficiários do Prouni. http://dadosabertos.mec.gov.br/prouni/item/124-bolsas-e-perfil-2020

[2] 2021. Dados da campanha de vacinação contra COVID-19. https://opendatasus.saude.gov.br/tl/dataset/covid-19-vacinacao

[3] 2021. Dados dos agendamentos para vacinação contra COVID-19 do Ceará. https://coronavirus.fortaleza.ce.gov.br/lista-vacinacao-d1.html

[4] John Corkett. [n. d.]. Data Anonymisation and L-Diversity. Retrieved May 12, 2019 from https://informationwithinsight.com/2019/03/12/data-anonymisation-and-l-diversity/

[5] Brasil. Superior Tribunal de Justiça (STJ). [n. d.]. LGPD - Lei geral de proteção de dados. Retrieved Jan 9, 2017 from https://www.stj.jus.br/sites/portalp/Leis-e-normas/lei-geral-de-protecao-de-dados-pessoais-lgpd

[6] Khaled Emam, Ann Brown, Philip Abdelmalik, Angelica Neisa, Mark Walker, Jim Bottomley, and Tyson Roffey. 2010. A method for managing re-identification risk from small geographic areas in Canada. *BMC medical informatics and decision making* 10 (04 2010), 18. https://doi.org/10.1186/1472-6947-10-18

[7] Philippe Golle. 2006. Revisiting the Uniqueness of Simple Demographics in the US Population. In *Proceedings of the 5th ACM Workshop on Privacy in Electronic Society (WPES '06)*. Association for Computing Machinery, New York, NY, USA, 77–80. https://doi.org/10.1145/1179601.1179615

[8] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. 2006. L-diversity: privacy beyond k-anonymity. In *22nd International Conference on Data Engineering (ICDE'06)*. 24–24. https://doi.org/10.1109/ICDE.2006.1

[9] Fabian Prasser and Contributors. 2012–2020. ARX (C). https://arx.deidentifier.org/.

[10] Latanya Sweeney. 2002. K-Anonymity: A Model for Protecting Privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* 10, 5 (oct 2002), 557–570. https://doi.org/10.1142/S0218488502001648

[11] Anderson F. V. Vidal. 2021. Script de ataque aos dados da campanha de vacinação da Paraíba contra a COVID-19. https://github.com/AndersonVidal/TCC-L-Diversidade-Dados-Vacinacao/blob/main/AtaquesProuni.ipynb.

[12] Anderson F. V. Vidal. 2021. Script de ataque aos dados da campanha de vacinação do Ceará contra a COVID-19. https://github.com/AndersonVidal/TCC-L-Diversidade-Dados-Vacinacao/blob/main/AntaqueAgendamentos.ipynb.

[13] Anderson F. V. Vidal. 2021. Script de conversão de PDF's em arquivo CSV. https://github.com/AndersonVidal/TCC-L-Diversidade-Dados-Vacinacao/blob/main/scripts/leitorPDF.py.

[14] Wantong Zheng, Zhongyue Wang, Tongtong Lv, Yong Ma, and Chunfu Jia. 2018. K-Anonymity Algorithm Based on Improved Clustering. 462–476. https://doi.org/10.1007/978-3-030-05054-2_36