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

ARTHUR FERNANDES DE ANDRADE

UM ESTUDO COMPARATIVO DE MECANISMOS DE PRIVACIDADE DIFERENCIAL APLICADO À RESOLUÇÃO DE ENTIDADES COM GARANTIAS DE PRIVACIDADE

CAMPINA GRANDE - PB

2023

ARTHUR FERNANDES DE ANDRADE

UM ESTUDO COMPARATIVO DE MECANISMOS DE PRIVACIDADE DIFERENCIAL APLICADO À RESOLUÇÃO DE ENTIDADES COM GARANTIAS DE PRIVACIDADE

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 em Ciência da Computação.

Orientador: Professor Dr. Carlos Eduardo Santos Pires.

CAMPINA GRANDE - PB

2023

ARTHUR FERNANDES DE ANDRADE

UM ESTUDO COMPARATIVO DE MECANISMOS DE PRIVACIDADE DIFERENCIAL APLICADO À RESOLUÇÃO DE ENTIDADES COM GARANTIAS DE PRIVACIDADE

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 em Ciência da Computação.

BANCA EXAMINADORA:

Professor Dr. Carlos Eduardo Santos Pires

Orientador – UASC/CEEI/UFCG

Professor Dr. Reinaldo Cezar de Morais Gomes

Examinador – UASC/CEEI/UFCG

Professor Tiago Lima Massoni

Professor da Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em: 14 de fevereiro de 2023.

CAMPINA GRANDE - PB

ABSTRACT

The progress of technology has generated an evolution in the popularity of services that use data exchange, allowing unknown or unreliable computers to keep a large amount of information about individuals based on their data provided. As a result, maintaining user privacy while ensuring service quality is a complex dilemma that has received attention in recent years. The integration of these various databases can occur through Entity Resolution (ER), however in many situations these data are private, something not supported by RE. Then comes the Privacy Entity Resolution (REGP) with the same purpose as the RE, but with the addition of data privacy support. One of the data protection techniques used in REGP is known as Differential Privacy (PD) which consists of using mechanisms to add noise to records. This work proposes to evaluate data privacy through PD mechanisms applied to REGP.

UM ESTUDO COMPARATIVO DE MECANISMOS DE PRIVACIDADE DIFERENCIAL APLICADO À RESOLUÇÃO DE ENTIDADES COM GARANTIAS DE PRIVACIDADE

Arthur Fernandes de Andrade
Universidade Federal de Campina Grande
Campina Grande, Paraíba, Brasil
arthur.andrade@ccc.ufcg.edu.br

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

RESUMO

O progresso da tecnologia tem gerado uma evolução da popularidade de serviços que utilizam a troca de dados, consentindo que computadores desconhecidos ou que não sejam confiáveis mantenham uma grande quantidade de informações dos indivíduos a partir de seus dados fornecidos. Como resultado, manter a privacidade dos usuários e ao mesmo tempo garantir a qualidade dos serviços é um dilema complexo que tem recebido atenção nos últimos anos. A integração dessas diversas bases de dados pode ocorrer por meio da Resolução de Entidades (RE), entretanto em muitas situações esses dados são de caráter privado, algo não suportado pela RE. Surge então a Resolução de Entidade com Garantia de Privacidade (REGP) com o mesmo propósito da RE, mas com a adição de suporte à privacidade dos dados. Uma das técnicas de proteção de dados utilizadas na REGP é conhecida como a Privacidade Diferencial (PD) que consiste em usar mecanismos para adicionar ruído aos registros. Este trabalho propõe avaliar a privacidade de dados através de mecanismos de PD aplicados à REGP.

Palavras-Chave

Resolução de Entidades; Privacidade Diferencial; Mecanismos; Privacidade.

1. INTRODUÇÃO

Atualmente, diversas empresas e instituições reúnem e processam uma quantidade significativa de dados para substanciar conhecimento e participar de maneira positiva em decisões. De acordo com Vatsalan [6], o mundo tem testemunhado uma explosão no volume de dados coletados por organizações e indivíduos, onde muitos desses dados são sobre pessoas ou gerados por pessoas.

Com o avanço da tecnologia, a troca de informações ocorre o tempo todo, os dados possuem grande valor para diversos tipos de organizações, sejam elas de varejo, saúde, entre outras áreas. Por exemplo, uma empresa relacionada à área de saúde, que consegue identificar quais locais estão mais propícios ou não à ocorrência de uma doença, ou uma empresa de varejo que possui, como técnica, a realização de vendas de acordo com o perfil dos consumidores. Segundo Machado et al.[4], isso só se torna possível através da análise de dados privados de indivíduos, levando a sérios riscos de exibir os dados particulares dos indivíduos.

Diferentes sistemas de informação recolhem e guardam dados nos quais aspectos como privacidade e confidencialidade devem se levar em conta. São exemplos destes dados: dados de navegação (GPS), transações financeiras, prontuários médicos, entre outros. A informação é considerada um requisito básico para o progresso econômico juntamente com o trabalho, a matéria-prima e o capital, mas o que torna a informação essencial atualmente é a sua natureza digital [1].

Muitas organizações executam análises críticas de dados para detectar padrões ocultos e antecipar tendências futuras. Para que análises sejam feitas, é indispensável que os dados estejam à disposição para acesso, seja por meio de publicações ou serviços de consulta [4]. Entretanto, os dados publicamente disponíveis podem incluir informações que identificam exclusivamente os indivíduos, causando assim uma violação da privacidade.

Por meio destas violações de Privacidade, processos que garantem a segurança dos dados são realizados para diferentes tipos de organizações. Um destes processos é conhecido como a Resolução de Entidades com Garantia de Privacidade (REGP), que visa identificar objetos similares em um conjunto de dados, garantindo que nenhuma informação de pessoa seja revelada a outra pessoa envolvida durante a realização da tarefa.

Este trabalho se justifica pela importância de fornecer informações para que titulares dos dados sejam capazes de

garantir o anonimato ao disponibilizar dados. A conservação do anonimato tornou-se um ponto de grande relevância, pois cada vez mais as organizações precisam fornecer dados ao público. Dessa forma, é de extrema importância pensar a respeito das seguintes perspectivas: é possível presumir a proteção de questões pessoais e ao mesmo tempo autorizar o acesso aos dados? O titular dos dados pode fornecer informações privadas de forma confidencial conservando a utilidade dos dados ?

Uma forma de anonimizar dados pessoais é com a utilização da técnica de Privacidade Diferencial (PD). Com a adição de ruídos aos dados, é possível gerar informações úteis com o conjunto de dados e ao mesmo tempo atrapalhar a identificação desses dados. O modelo de Privacidade Diferencial surgiu nas últimas décadas como um modelo matemático rigoroso que provê fortes garantias de privacidade [3]. A Privacidade Diferencial assegura que qualquer sequência de respostas de consultas é igualmente possível de ocorrer independente da presença, ou ausência, de qualquer item (por exemplo, um indivíduo) no conjunto de dados [3].

O objetivo geral deste trabalho é analisar o comportamento dos dados ao adicionar ruído a um conjunto de dados, utilizando conceitos que dizem respeito à PD em termos de proteção dos dados. Neste processo, foi utilizado PD para proteger as informações de similaridade com valores que variam entre 0 e 1 quando utilizados na REGP.

Este artigo está estruturado da seguinte forma: na seção 2, se fornece uma explicação teórica sobre PD, REGP, Mecanismo de Laplace e Mecanismo Gaussiano, para auxiliar no entendimento das seções seguintes. Na seção 3, é apresentada a metodologia utilizada durante o estudo. Na seção 4, são discutidos os resultados obtidos e a análise realizada sobre os dados. Finalmente, na seção 5, são apresentadas as conclusões e trabalhos futuros.

2. FUNDAMENTAÇÃO TEÓRICA

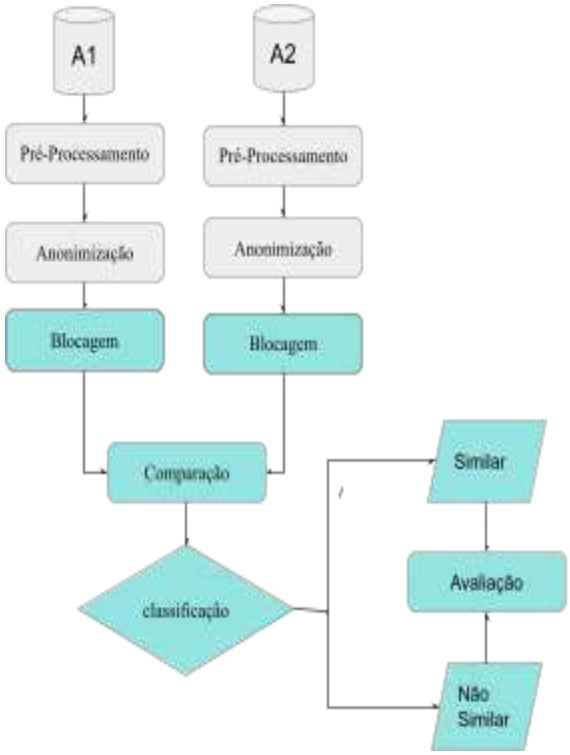
Esta seção apresenta os conceitos referentes à proteção de dados e aos mecanismos relacionados ao contexto da privacidade diferencial.

2.1 Resolução de Entidades com Garantia de Privacidade

Para a garantia de segurança, os algoritmos de REGP usam técnicas para anonimizar os dados e tornar mais difícil a inferência de informações por terceiros. A Figura 1 ilustra a tarefa de REGP com os dados. As etapas mostradas na cor verde utilizam-se dos dados originais (também chamados de planos ou não anonimizados) enquanto que as etapas mostradas na cor cinza usam os dados anonimizados.

.

Figura 1. Fluxo da tarefa de REGP



Fonte: adaptado de Vatsalan et. al. 2013.

A etapa de Pré-processamento tem a responsabilidade de corrigir os dados e resolver os problemas de diferença, representando os dados em um formato que ocasione a comparação entre entidades [2].

A etapa de Anonimização tem a finalidade de mascarar os dados de maneira que possam ser usados nas etapas seguintes [6].

A etapa de Blocagem busca reduzir o número de comparações entre entidades a serem realizadas na etapa posterior (etapa de comparação).

Na etapa de Comparação, funções de similaridade são aplicadas aos dados anonimizados, essas funções produzem valores que quantificam o grau de similaridade entre as duas entidades. Na etapa de Classificação, com base nos valores calculados na etapa de comparação, é possível determinar quais pares de entidades são realmente similares ou não. Uma vez concluída a etapa de Classificação, é revelado aos participantes da tarefa REGP o número de entidades que foram classificadas como similares.

A última etapa da tarefa REGP é a de avaliação, que implica em investigar o desempenho, a qualidade e a privacidade da REGP.

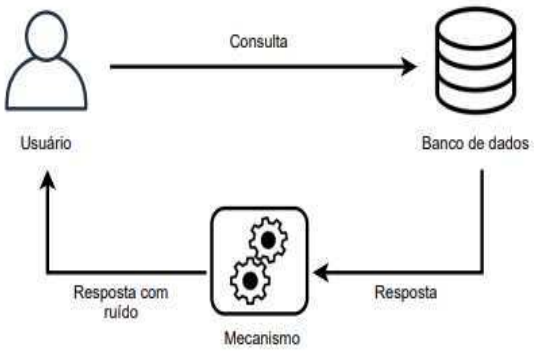
Como nem sempre é possível obter soluções exatas para problemas de REGP, a maioria dos algoritmos de REGP atualmente disponíveis usam técnicas de otimização para a obtenção de bons resultados. Uma dessas técnicas é conhecida

como a privacidade diferencial, que visa melhorar a privacidade de qualquer dado que seja.

2.2 Privacidade Diferencial

O modelo de privacidade diferencial procura sustentar a preservação de privacidade na publicação de resultados de consultas. Seu propósito é impedir que o conhecimento adversário de um atacante aumente a probabilidade de expor os indivíduos do conjunto de dados, evitando ataques probabilísticos. Para esta finalidade, as respostas destas consultas são perturbadas, com a adição de ruído, como forma de garantir a privacidade dos indivíduos, como mostrado na Figura 2, onde se mostra a relação dos dados com o mecanismo.

Figura 2. Ambiente interativo no modelo de Privacidade Diferencial



Fonte: Javam C. Machado et al. 2019.

A PD é um modelo semântico, cujo objetivo é assegurar a utilidade dos dados ao mesmo tempo que fornece proteção contra ataques de conhecimento prévio. Segundo Near e Abuah [5], a PD é uma noção formal de privacidade. Uma função que atende a privacidade diferencial é regularmente chamada de mecanismo.

O mecanismo é o responsável por adicionar ruído controlando a resposta da consulta com o intuito de assegurar ε-Privacidade Diferencial [5]. A quantidade de ruído que é indispensável depende do tipo de consulta foi aplicada sobre o conjunto de dados D .

No decorrer do processo de privacidade, algumas propriedades de mecanismos são relevantes para se obter dados úteis que satisfaçam a privacidade diferencial e garantam que esses mesmos algoritmos fornecem respostas precisas. São eles a composição sequencial, composição paralela e o pós processamento.

Na propriedade de composição sequencial é a própria composição que determina um limite no custo total de privacidade de liberar múltiplos resultados de mecanismos diferencialmente privados nos mesmos dados de entrada.

A composição sequencial é uma propriedade vital da privacidade diferencial porque permite o projeto de algoritmos que consultamos dados mais de uma vez.

A composição paralela se baseia na ideia de dividir seu conjunto de dados em pedaços separados e executar um mecanismo privado em cada pedaço separadamente.

No caso do pós-processamento, a propriedade tem como significado a segurança de realização de cálculos arbitrários na saída de um mecanismo privado, não havendo perigo de reverter a proteção de privacidade que o mecanismo forneceu fazendo assim que possa acontecer a redução do ruído além de melhorar o sinal na saída do mecanismo.

De maneira geral, a quantidade de ruído necessária para garantir a privacidade diferencial para uma determinada consulta depende do fator s (sensibilidade), que é o valor máximo de ruído que pode se adicionar ao dado [5]. Desta maneira, obtendo-se um resultado final no que se diz respeito à utilização de mecanismos exponenciais de ruído apresentados, verifica-se a utilização desses mesmos mecanismos para proteção de um conjunto de dados podendo avaliar a qualidade e privacidade. Por meio disto um meio para alcançar a privacidade diferencial em consultas sobre dados numéricos que retornam valores agregados, é utilizando o mecanismo de Laplace, onde o principal desafio é adicionar ruído suficiente para satisfazer a definição privacidade diferencial de modo que a resposta não se torne muito barulhenta.

2.3 Mecanismo Laplaciano

O mecanismo de Laplace é um dos algoritmos com menor complexidade para adicionar ruído e alcançar a PD. De acordo com o mecanismo de Laplace, para uma função $f(x)$ que retorna um número, a seguinte definição de $F(x)$ satisfaz a privacidade ϵ-diferencial: onde s é a sensibilidade de f e Lap(S) denota a amostragem da distribuição de Laplace com centro 0 e escala S.

$$F(x) = f(x) + Lap\left(\frac{S}{\epsilon}\right) \quad (1)$$

Para melhor entendimento da utilização do mecanismo de Laplace, a Figura 3 mostra uma tabela contendo quatro registros referentes a um conjunto de dados hipotéticos de clientes de agência bancária, além da quantidade de contas correntes que cada cliente possui.

Figura 3. Exemplo de conjunto de dados originais contendo a quantidade de contas correntes de cada indivíduo.

ID	Cliente	Nº Contas
1	Maria	4
2	Jose	2
3	Luis	7
4	Carlos	1

Fonte: adaptado de Javam C. Machado et al. 2019.

Assuma a consulta f que devolve o total de contas de todos os clientes da base de dados. Primeiramente, o que deve ser feito é o cálculo da sensibilidade da função f sobre os dados. Para isso, é calculado f para cada conjunto de dados vizinho. A resposta real é 11, pois é a soma do número de contas existentes. A Figura 4 mostra os conjuntos de dados vizinhos gerados de acordo com os dados originais e suas respostas à consulta f.

Figura 4. Conjuntos de dados vizinhos gerados a partir do conjunto de dados originais e suas respostas à consulta f(soma).

ID	Cliente	Nº de Contas
2	Jose	2
3	Luis	7
4	Carlos	1

F(D1) = 2 + 7 +1 =10

ID	Cliente	Nº de Contas
1	Maria	4
3	Luis	7
4	Carlos	1

F(D2) = 4 + 7 +1 =12

ID	Cliente	Nº de Contas
1	Maria	4
2	Jose	2
4	Carlos	1

F(D3) = 4 + 2 +1 = 7

ID	Cliente	Nº de Contas
1	Maria	4
2	Jose	2
3	Luis	7

F(D4) = 4 + 2 +7 =13

Fonte: adaptado de Javam C. Machado et al. 2019.

Dessa maneira, a sensibilidade é dada pela mudança máxima que a falta de um cliente provoca no resultado da consulta. Essa mudança é alcançada quando se retira o registro de id = 4, cuja diferença máxima é de 7. E finalmente, o ruído que será introduzido para se obter ao modelo de Privacidade Diferencial, utilizando o mecanismo de Laplace, deve ser igual a Laplace(0, $\frac{7}{\epsilon}$)

Outro mecanismo que também adiciona ruído de forma aleatória é o Gaussiano, diferenciando-se do mecanismo de Laplace pela maneira que o ruído é adicionado.

2.4 Mecanismo Gaussiano

O Mecanismo Gaussiano adiciona ruído Gaussiano ao invés do ruído de Laplace como uma alternativa ao Mecanismo Laplace [5]. O mecanismo gaussiano satisfaz a privacidade diferencial (ϵ, δ), mas não a privacidade diferencial ϵ pura. De acordo com o mecanismo gaussiano, a seguinte definição de F(x) satisfaz a privacidade diferencial :

$$F(x) = f(x) + \mathcal{N}(\sigma^2) \quad (2)$$

, onde $\mathcal{N}(\sigma^2)$ denota a simplificação da distribuição gaussiana com mediana nula e σ^2 como variância. onde σ^2 é dado por:

$$\frac{2s^2\, log(1.25/\delta\ )}{\epsilon^2} \quad (3)$$

, em que s denota a sensibilidade de F, δ é a probabilidade de falha e o valor de epsilon denotado como ϵ é a garantia de privacidade.

3. METODOLOGIA

A metodologia utilizada para tratar o problema referente a proteção em um conjunto de dados utilizando mecanismos da privacidade diferencial será detalhada nesta seção.

3.1 Planejamento experimental

A presente pesquisa foi realizada através do Google Colab usando implementação na linguagem de programação Python, com intuito de adicionar ruído Laplaciano e Gaussiano a um conjunto de dados, para manter a privacidade.

Inicialmente, foi investigado o conceito que se diz respeito à anonimização dos dados, similaridade e garantia de privacidade referente à ideia da privacidade diferencial (PD) para que pudesse ser feito um estudo mais preciso sobre a proteção dos dados.

Depois de toda teoria consolidada, foi realizada a implementação com embasamento no mecanismo Gaussiano onde foi adicionado ruído com dados numéricos de ponto flutuante que variam entre 0 e valores aproximados de 1 para ϵ (garantia de privacidade) e s(sensibilidade), durante o processo de desenvolvimento foram gerados gráficos com a amostra dos dados antes e depois de se adicionar o ruído, onde cada ponto no eixo Y representam os valores de registro que foram estudados, e cada ponto no eixo X representam as posições que estes registros se encontram no conjunto de dados, como mostrado na Figura 7.

Com o estudo dos dados anonimizados para ambos os mecanismos foi realizada uma comparação e obtenção dos resultados por meio da análise dos gráficos que foram plotados com a adição de ruído Gaussiano e Laplaciano .

3.2 Tratamento dos Dados

Para realização do estudo, foi exportado um conjunto de dados contendo 8 atributos e 23.508 registros contendo valores de similaridade entre entidades de uma tarefa de REGP com valores que variam de 0 a 1, como mostrado na figura 5.

Figura 5. Amostra dos registros

	id1	id2	dice	jaccard	overlap	hamming	entropy	is_match
0	1	1	1.000000	1.000000	1.000000	1.00	1.000000	0
1	1	2	1.000000	1.000000	1.000000	1.00	1.000000	1
2	1	24	0.361111	0.220339	0.382353	0.79	0.951444	0
3	1	25	0.361111	0.220339	0.382353	0.79	0.951444	0
4	1	28	0.393939	0.245283	0.406250	0.81	0.792928	0
...	...	...	...	...	...	...	...	...
23503	727	727	1.000000	1.000000	1.000000	1.00	1.000000	0
23504	728	728	1.000000	1.000000	1.000000	1.00	1.000000	0
23505	741	741	1.000000	1.000000	1.000000	1.00	1.000000	0
23506	755	755	1.000000	1.000000	1.000000	1.00	1.000000	0
23507	778	778	1.000000	1.000000	1.000000	1.00	1.000000	0

23508 rows × 8 columns

Fonte: Do autor (2023).

Baseado nas referências teóricas desta pesquisa para cada mecanismo que foi testado foram escolhidos valores de s e ϵ que fossem menores e maiores do que 1 com o intuito de verificar como o ruído seria adicionado aos dados estudados.

Em s foram testados 3 valores , que são eles: 0.025 quando for menor que 1, 1.050 como um valor aproximado de 1 e 3 como uma valor maior que 1. No caso de ϵ foram testados 4 valores: 0.025 e 0.78 quando for menor que 1, 1.50 para um valor aproximado de 1 e 2.50 quando o valor de ϵ for maior 1.

Com os testes realizados, os melhores valores para os mecanismos foram definidos como 0.025 para s, pois se percebeu que quando o valor da sensibilidade ultrapassa 1 independente do valor de ϵ os dados não eram mascarados da melhor forma e 0.78 para ϵ por notar que valores abaixo de 1 geram um melhor ruído nos dados

.

4. RESULTADOS E DISCUSSÃO

Os valores interligados ao mecanismo de Laplace e Gaussiano utilizados nesta pesquisa foram modificados para verificar o comportamento das respostas das consultas, para se obter uma conclusão mais precisa referente a comparação e resolução dos dados utilizando o conceito dos mecanismos estudados.

As análises dos resultados foram feitas a partir de gráficos que mostram os dados com ruído em laranja sobrepondo os dados originais mostrado em azul.

Para poder ser feita uma análise do comportamento dos dados foram utilizados nos mecanismos Laplaciano e Gaussiano uma média de ruído centrada em 0, empiricamente o valor de s foi atribuído sendo igual a 0.025 e o valor de ϵ foi definido como 0.78 para se manter uma privacidade aceitável de maneira que não afetasse agressivamente os dados.

Foi observado que, quanto maior o valor de ϵ, menor será o desvio padrão. Com o desvio padrão sendo reduzido, o ruído se aproxima mais da média fazendo com que o ruído altere menos os dados, mas diminua a privacidade.

4.1 Resultados utilizando o Mecanismo Gaussiano

No resultado do experimento mostrado na Figura 8, percebe-se que a cada posição do registro é adicionado o valor do registro mais um valor de ruído, fazendo com que o valor do registro varie de maneira positiva ou negativa no eixo Y.

Nesse sentido percebe-se que quanto maior a distância entre o valor do dado original e o valor do dado com o ruído, maior é o ruído que está sendo adicionado.

Por meio do mecanismo gaussiano, se conseguiu adicionar privacidade no experimento, pois ao observarmos os dados com ruído representado em laranja na figura 8, percebemos que se comparados aos dados originais houve mudança em relação aos valores dos registros da Figura 7. Isso se torna perceptível pois o comportamento dos dados com ruído no gráfico da Figura 8 é diferente dos dados originais do experimento.

4.2 Resultados utilizando o Mecanismo Laplaciano

Um ruído Laplaciano foi adicionado ao experimento utilizando os mesmos dados mostrados na Figura 9, com o intuito de melhorar o entendimento do comportamento dos mecanismos estudados.

Durante os experimentos, ao utilizar o mecanismo de Laplace percebeu-se que a privacidade dos dados estudados foi atendida, como mostra a Figura 9. Pois o comportamento dos dados com ruído exibiu diferença ao comportamento dos dados originais apresentados na Figura 7, implicando que os registros originais foram mascarados com o valor do registro mais o ruído.

4.3 Comparações dos resultado dos mecanismos utilizados

Depois de verificar os resultados por meio dos mecanismos estudados, foi feita uma comparação para verificar qual deles obteve melhor desempenho ao mascarar os dados.

Foi observado quantitativamente os ruídos que foram adicionados em ambos os mecanismos, como mostrado nas figuras 6, onde

cada valor está sendo mostrado, acompanhado do valor de ruído que foi adicionado.

Figura 6. Resultados dos valores com ruído Gaussiano e Laplaciano.(C) Ruído Gaussiano (D) Ruído Laplaciano.

(C) Ruido Gaussiano

	Dados orgniais	Dados com Ruido	ruido
0	1.000000	1.052572	0.052572
1	1.000000	0.990366	-0.009634
2	0.361111	0.331980	-0.029131
3	0.361111	0.309283	-0.051828
4	0.393939	0.380460	-0.013480
...	...	...	...
23503	1.000000	1.098893	0.098893
23504	1.000000	0.972966	-0.027034
23505	1.000000	1.001295	0.001295
23506	1.000000	0.976865	-0.023135
23507	1.000000	1.027871	0.027871

[23508 rows x 3 columns]

(D) Ruido Laplaciano

	Dados orgniais	Dados com Ruido	ruido
0	1.000000	0.992961	-0.007039
1	1.000000	0.998301	-0.001699
2	0.361111	0.337334	-0.023777
3	0.361111	0.348732	-0.012379
4	0.393939	0.383287	-0.010652
...	...	...	...
23503	1.000000	0.967615	-0.032385
23504	1.000000	0.979647	-0.020353
23505	1.000000	0.975573	-0.024427
23506	1.000000	1.082411	0.082411
23507	1.000000	1.013985	0.013985

[23508 rows x 3 columns]

Fonte: Do autor (2023).

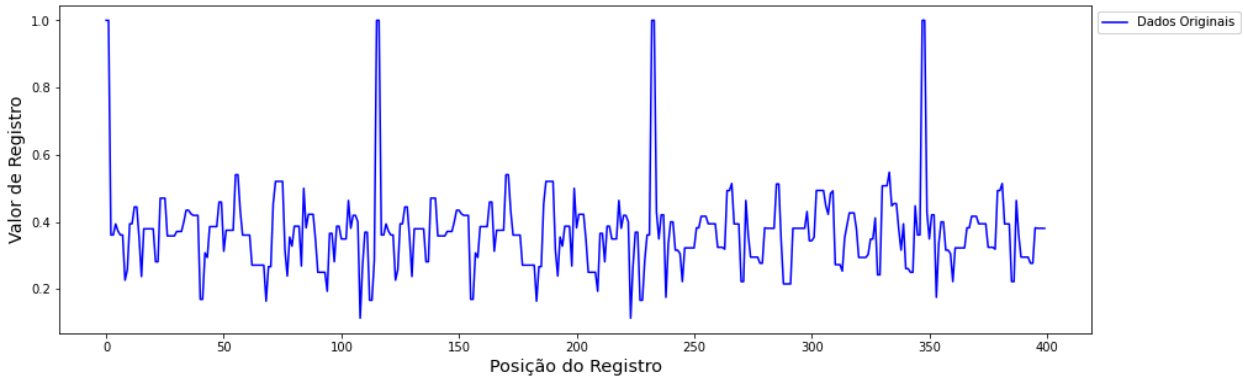
Com a análise que foi realizada , percebeu-se que quanto maior a diferença do valor do dado com ruído em relação ao valor do dado original maior é o ruído que vai ser adicionado. Em virtude disso, os dados com ruído mostrados em C na Figura 6 possuem uma maior diferença dos dados originais quando comparado com D.

Isso pode ser observado na Figura 10, que mostra os resultados obtidos durante os estudos, aplicando valores iguais para ϵ e s no mesmo conjunto de dados.

Ao comparar os gráficos A e B nota-se que o mecanismo Gaussiano gerou mais ruído e consequentemente maior privacidade.

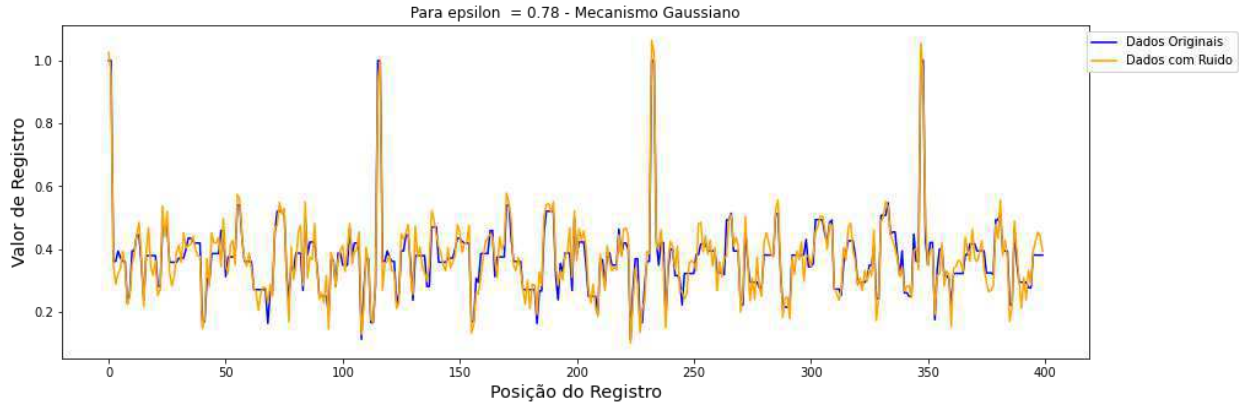
Por meio disto percebeu-se que o Mecanismo Gaussiano apresentou melhores resultados se comparado com o mecanismo Laplaciano, por mostrar dados com ruído que são menos similares aos dados originais.

Figura 7.. Dados Originais



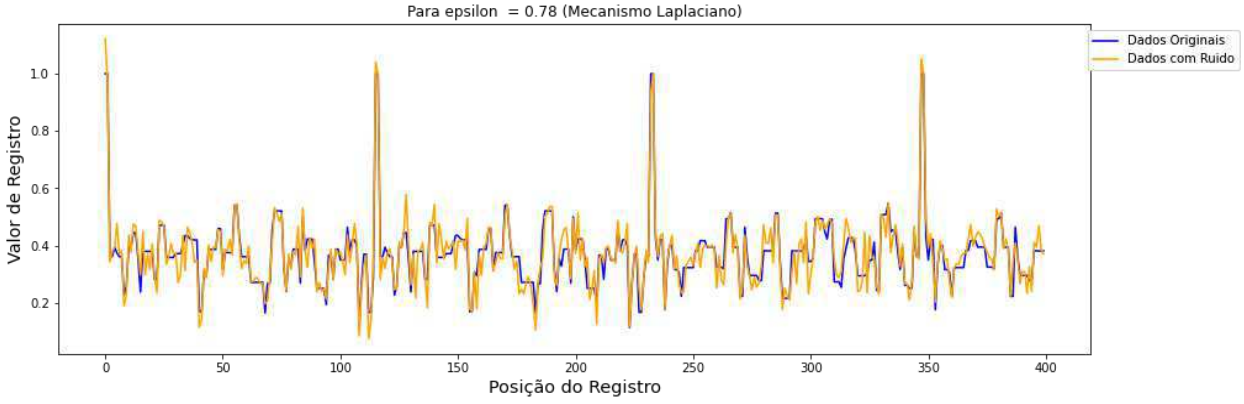
Fonte: Do autor (2023).

Figura 8. Resultado do ruído usando o Mecanismo Gaussiano nos 400 primeiros registros.



Fonte: Do autor (2023).

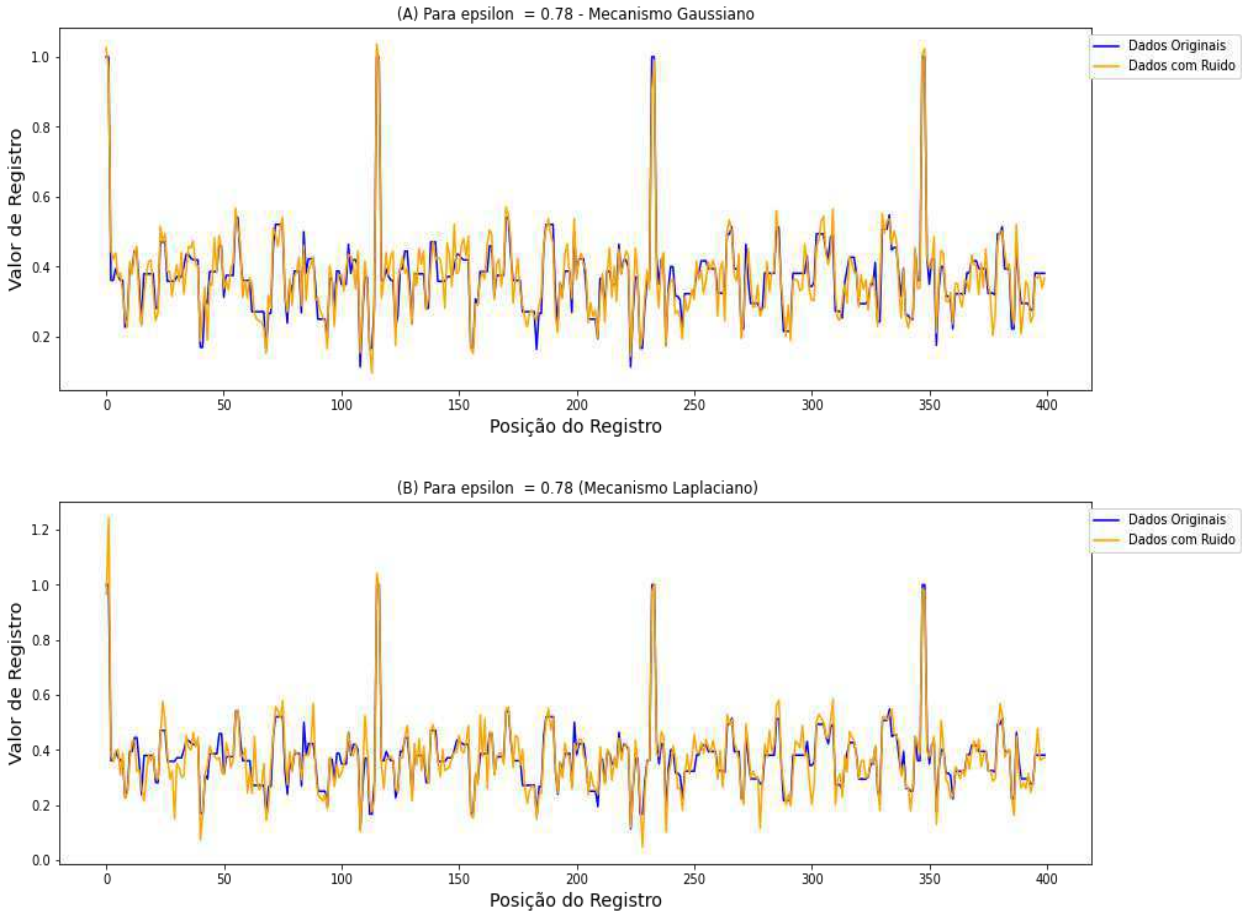
Figura 9. Resultado do ruído usando o Mecanismo de Laplace nos 400 primeiros registros.



Fonte:Do autor (2023).

Figura 10. Comparação dos resultados utilizando Mecanismo Gaussiano e Laplaciano nos 400 primeiros registros.

(A) Dados utilizando mecanismo Gaussiano (B) Dados utilizando mecanismo Laplaciano.



Fonte: Do autor (2023).

5. CONCLUSÕES E TRABALHOS FUTUROS

Diversas organizações precisam divulgar dados para a sociedade ao mesmo tempo em que protegem a privacidade dos seus usuários. Mesmo quando utilizamos técnicas elaboradas como a Privacidade Diferencial, pode não ser simples definir qual o melhor mecanismo a ser usado em um conjunto de dados. Neste artigo foi avaliada uma abordagem para a Resolução de proteção de dados utilizando técnicas de privacidade diferencial, com uso de mecanismos. A partir dos resultados obtidos, foi possível inferir qual dos mecanismos mantém o melhor nível de utilidade para anonimizar os dados. Embora o Mecanismo Gaussiano mostrou ter mais utilidade ao anonimizar os dados estudados, percebeu-se que o mecanismo de Laplace atendeu as expectativas, e manteve a privacidade dos dados. De maneira geral, a abordagem se mostrou compatível com a resolução do problema para o conjunto de dados avaliados. Como sugestões para trabalhos futuros, espera-se investigar mais a respeito da utilidade dos dados e ver o impacto da privacidade diferencial versus a utilidade.

6. REFERÊNCIAS

[1] CAPURRO, Rafael; HJORLAND, Birger. The Concept of Information. Theorizing Information and Information Use. Annual Review of informatiocn Science and Technology. v. 37, cap. 8, p.343-411, 2003.

[2] CHRISTEN, Peter . Data Matching. Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection, 2012.

[3] MENDONCA, André. Uma abordagem de privacidade diferencial para dados correlacionados utilizando técnicas de agrupamento. Machado, Javam de Castro. Programa de Pós Graduação em Ciências da Computação, Fortaleza. 2018 Disponível em: http://www.repositorio.ufc.br/handle/riufc/38796

[4] MACHADO, Javam, NETO, Eduardo, FILHO, Emanuel. Técnicas de Privacidade de Dados de Localização: Tópicos em Gerenciamento de Dados e Informações 2019. cap 1 , 2019.

[5] NEAR, Joseph, ABUAH,Chiké. Programming Differential Privacy: programming Differential Privacy 2022 . cap 2- 7, 2022.

[6] VATSALAN, Dinusha, et al.. A taxonomy of privacy preserving record linkage techniques. Information Systems, 2013.

Sobre autor:

Arthur Fernandes de Andrade é graduando em Ciência da Computação e atualmente atua como monitor na disciplina de lógica para computação e professor de Matemática particular.

Carlos Eduardo Santos Pires. Professor Dr. na UFCG. Orientador.