Universidade Estadual de Londrina

RENAN RICOLDI FRÓIS PEDRO

PRIVACIDADE DIFERENCIAL PARA PROTEGER DADOS SOBRE UTILIZAÇÃO DE BICICLETAS COMPARTILHADAS

LONDRINA
2022

RENAN RICOLDI FRÓIS PEDRO

PRIVACIDADE DIFERENCIAL PARA PROTEGER DADOS SOBRE UTILIZAÇÃO DE BICICLETAS COMPARTILHADAS

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Ciência da Computação da Universidade Estadual de Londrina para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Bruno Bogaz Zarpelão

LONDRINA

2022

Ficha de identificação da obra elaborada pelo autor, através do Programa de Geração Automática do Sistema de Bibliotecas da UEL

P372p Pedro, Renan Ricoldi Fróis.
Privacidade diferencial para proteger dados sobre utilização de bicicletas compartilhadas / Renan Ricoldi Fróis Pedro. - Londrina, 2022.
40 f. : il.

Orientador: Bruno Bogaz Zarpelão.
Trabalho de Conclusão de Curso (Graduação em Ciência da Computação) - Universidade Estadual de Londrina, Centro de Ciências Exatas, Graduação em Ciência da Computação, 2022.
Inclui bibliografia.

1. Privacidade diferencial - TCC. 2. Proteção de dados - TCC. I. Zarpelão, Bruno Bogaz. II. Universidade Estadual de Londrina. Centro de Ciências Exatas. Graduação em Ciência da Computação. III. Título.

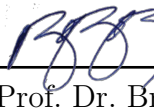
CDU 519

RENAN RICOLDI FRÓIS PEDRO

PRIVACIDADE DIFERENCIAL PARA PROTEGER DADOS SOBRE UTILIZAÇÃO DE BICICLETAS COMPARTILHADAS

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Ciência da Computação da Universidade Estadual de Londrina para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA



Orientador: Prof. Dr. Bruno Bogaz Zarpelão
Universidade Estadual de Londrina

Prof. Dr. Guilherme Pina Cardim
Universidade Estadual de Londrina

Fernando Henrique Yoshiaki Nakagawa
Universidade Estadual de Londrina

Londrina, 01 de junho de 2022.

AGRADECIMENTOS

Começo agradecendo os meus professores, que me concederam o conhecimento, em especial o professor Bruno Bogaz Zarpelão, que deu todo o suporte necessário para o desenvolvimento desse trabalho. Também agradeço a minha família, que sempre batalhou para que eu pudesse ter a melhor educação. Agradeço os meus amigos, que durante o percurso da faculdade me motivaram para não desistir, em especial o Alan Leiser e o Fernando Morgado, que se mostraram leais nesse caminho. Por fim, agradeço a minha namorada Giulia Beatriz Gasparini, que me ajudou em todos os momentos difíceis da minha graduação.

PEDRO, R. R. F.. **Privacidade diferencial para proteger dados sobre utilização de bicicletas compartilhadas**. 2022. 40f. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) – Universidade Estadual de Londrina, Londrina, 2022.

RESUMO

A privacidade pode ser definida como o controle de um indivíduo sobre o uso correto dos seus dados. Atualmente, vários serviços *on-line* têm acesso aos dados de seus usuários, portanto, entende-se como importante que a privacidade desses dados seja garantida. A privacidade diferencial propõe uma forma de aprender propriedades de um conjunto de dados sem ferir a privacidade dos usuários. Através de ruído, consegue impedir a identificação de um dado, garantindo a privacidade diferencial dele, porém diminuindo a utilidade do dado para o aprendizado. O objetivo do trabalho é aplicar duas técnicas de privacidade diferencial, uma no modo central e a outra no modo local, para entender o comportamento de cada uma em relação à privacidade e utilidade dos resultados. Para isso, foram feitos testes com cada técnica usando diferentes níveis de privacidade. Os testes utilizaram conjuntos de dados de aluguel de bicicletas compartilhadas em Nova Iorque e, com os resultados obtidos, foram analisadas métricas de erro absoluto e relativo.

Palavras-chave: preservação de privacidade; controle de inferência; segurança da informação; mineração de dados.

PEDRO, R. R. F.. **Differential privacy for protecting bike sharing use data**. 2022. 40p. Final Project (Bachelor of Science in Computer Science) – State University of Londrina, Londrina, 2022.

ABSTRACT

Privacy may be defined as the control a person has over the correct usage of its data. Nowadays, plenty of on-line services have access to user data, thus, it is understood that user data privacy must be guaranteed. Differential privacy proposes a way of learning properties of a dataset, without jeopardize user's privacy. By applying noise, it can prevent data identification, assuring its differential privacy, though loosing data usability for learning. The aim of this paper is to apply two differential privacy techiniques, one using the central mode and the other the local mode, in order to understand their behavior on privacy and usability of their results. To this end, tests were made using each technique with different levels of privacy. The tests used a dataset of bike sharing in New York, and with the obtained results, absolute and relative error metrics were analised.

Keywords: privacy preserving; inference control; information security; data mining.

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

DP	Differential Privacy
LDP	Local Differential Privacy
RR	Randomized Response
MAE	Mean Absolute Error

SUMÁRIO

1 INTRODUÇÃO

Com o acréscimo do tempo que as pessoas passam conectadas à Internet e dispositivos móveis, o número de dados coletados também aumenta. Informações pessoais como características físicas do usuário, histórico de buscas, compras e localização em tempo real são obtidas para aprimorar serviços e produtos em diversas áreas, incluindo planejamento de tráfego, publicidade direcionada e saúde pública [2] [3].

O registro de dados de localização de um indivíduo consiste em um conjunto de posições relacionadas a ele. Obtendo uma coleção de registros de localização é possível extrair informações importantes e sensíveis das pessoas. Por exemplo, ao observar a frequência com que alguém visita clínicas e consultórios médicos, pode-se deduzir seu estado de saúde [4]. Assim, apesar da utilidade desses dados, existe um risco de privacidade para os usuários e, da mesma forma, é um problema para o coletor dos dados, especialmente em vazamentos de grandes coleções, que podem causar perda de usuários e aplicação de ações judiciais, como nos incidentes com a Uber[1], Netflix[2] e Snapchat[3] [3].

De forma a resolver as preocupações com a privacidade, foi proposta a privacidade diferencial (DP - *differential privacy*), uma promessa de que o usuário não será prejudicado por fornecer dados para pesquisas e análises. Para isso, mecanismos aplicam ruído nos dados, que pode alterar ou não seus valores, e os tornam diferencialmente privados, o que permite que sejam usados para estudos mesmo contendo informações confidenciais [5].

A DP em seu modo padrão, o central, precisa que os dados dos usuários sejam compartilhados com um coletor, e disponibilizados para um curador, que é responsável por aplicar a DP nos dados e responder as análises necessárias sem revelá-los. No entanto, esse esquema necessita que o curador seja confiável, e isso pode fazer com que os proprietários dos dados não se sintam seguros ao compartilhar os dados reais com o coletor. Uma solução para este problema é a privacidade diferencial local (LDP - Local Differential Privacy), na qual cada usuário aplica ruídos nas próprias informações, e envia ao coletor apenas os dados já diferencialmente privados [1].

Neste trabalho, serão aplicadas duas técnicas de privacidade diferencial usando dados de aluguel de bicicletas compartilhadas. O objetivo é avaliar a relação entre privacidade e utilidade dos resultados diferencialmente privados, comparando com os resultados obtidos sobre os dados originais. A primeira técnica utiliza a privacidade diferencial central, e consegue ser útil mesmo aplicando maior nível de ruído aos dados originais, o que garante a privacidade diferencial, e ao mesmo tempo entrega um resultado parecido com

[1] <uber.com/en-CA/newsroom/2016-data-incident>
[2] <wired.com/2009/12/netflix-privacy-lawsuit>
[3] <techland.time.com/2014/01/01/hackers-reveal-4-6-million-snapchat-usernames-and-phone-numbers>

aquele dos valores reais. A segunda técnica, realizada com a privacidade diferencial local, evita a necessidade de compartilhar os dados reais com o coletor, porém, precisa que pouco ruído seja aplicado para que o resultado seja útil, o que pode diminuir o nível de privacidade

No Capítulo 2, é apresentado o conceito de privacidade diferencial, dois modos de aplicá-la, central e local, além das vantagens que cada modo possui. No Capítulo 3, são explicados os materiais e métodos usados, incluindo o conjunto de dados utilizado, os cálculos feitos para obter os resultados e os mecanismos para garantir a privacidade diferencial. No Capítulo 4, os resultados dos experimentos são apresentados e discutidos. Por fim, no Capítulo 5, as ideias do trabalho são concluídas.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Privacidade Diferencial

O avanço no número de usuários da Internet a torna cada vez mais importante para as atividades socio-econômicas cotidianas da humanidade. Os serviços digitais geram quantidades massivas de dados que normalmente são usados para aprimorar a experiência dos usuários. Esses dados podem ser informações sensíveis do usuário, como nome, idade e endereço. Mesmo quando não contém informações tão óbvias do usuário, ainda pode conter dados que ao serem combinados com informações auxiliares, revelem informações específicas dos usuários [5].

Essas grandes coleções de dados normalmente são salvas em bancos de dados, que constituem uma forma de salvar digitalmente um conjunto de dados de forma estruturada em um computador. Os dados são salvos através de registros que seguem a mesma estrutura das informações salvas, de forma que cada registro do banco de dados possua os mesmos tipos de dados, porém com informações referentes ao registro. Por exemplo, podemos ter um banco de dados com histórico de pessoas que alugaram bicicletas disponibilizadas em pontos de aluguel automatizados. Cada registro pode conter a identificação da bicicleta, ano de nascimento e gênero do usuário, além do ponto de locação e horário de retirada e devolução.

No entanto, essa grandes coleções de informações sensíveis dos usuários estão sujeitas à utilização e armazenamento inadequados, ferindo a privacidade deles. A privacidade pode ser definida de várias formas, as quais tipicamente englobam o controle dos usuários sobre a possibilidade do uso de seus dados para outros propósitos além dos originalmente informados [6].

Assim, cria-se uma balança onde a coleta de dados é muito importante para melhorar os serviços disponíveis online, porém podem ser usados para outras ações que prejudiquem o usuário. Por exemplo, o Facebook usa dados pessoais para apresentar um conteúdo direcionado a cada usuário, porém esses mesmos dados foram usados para tentar influenciar o voto dos usuários nas eleições nos Estados Unidos [7].

Dessa forma, para as empresas que coletam os dados garantirem a privacidade dos usuários, faz-se necessária a criação de leis e regulações que obrigam a ter um nível mínimo de proteção dos dados [7]. Infelizmente, ações jurídicas são parcialmente efetivas, gerando a necessidade de técnicas que permitam que analistas aprendam algo de dados sensíveis, sem revelar informações específicas dos usuários.

Esse objetivo impede que técnicas como criptografia sejam usadas, uma vez que

após criptografados, os dados não revelam nada. Além disso, abordagens que retiram as informações que identifiquem alguém, como nome e número de documento, não garantem que o usuário não será identificado [8]. Os *Linkage Attacks* (ataques de ligação) combinam os dados disponibilizados com informações auxiliares, como a idade de uma pessoa e o horário em que ela costuma alugar bicicletas, por exemplo, para identificar a quem pertence um registro do banco de dados, e suas informações de aluguel de bicicletas. Assim, para resolver esse problema, uma maneira de definir o nível de privacidade de um banco de dados foi desenvolvida [8].

A privacidade diferencial (DP - *differential privacy*), proposta por Dwork et al. [9], define uma forma de aprender propriedades de uma população como um todo, sem ferir a privacidade de cada indivíduo que a compõe. Mais precisamente, o fato de alguém possuir suas informações em um banco de dados usado com privacidade diferencial não deve aumentar de forma significativa o risco de prejudicar a sua privacidade se comparado a alguém cujas informações não estão no banco. Mesmo assim, se alguém quiser aprender informações úteis sobre a população, conseguirá. Desta forma, independente da presença ou ausência de um indivíduo na base de dados, o resultado do estudo não deve ser alterado. No entanto, DP não cria privacidade em dados onde ela já não existe, como quando existe um número pequeno de dados ou se o resultado revela algo sobre o grupo que alguém pertence, mas quantifica a perda de privacidade [5] [8].

Computacionalmente, a DP, no seu modo central, supõe que um coletor de dados confiável adquire informações sensíveis de um grande grupo de indivíduos para poder aprender fatos estatísticos sobre o grupo e apresentar ao público [5]. Com esse propósito, ela guarda as informações reais da população em um banco de dados, uma coleção de dados estruturada e salva digitalmente em um computador. Esse banco de dados é acessível apenas por um curador, o qual será o único responsável a responder perguntas sobre a população [1].

O coletor de dados tem como único trabalho colher os dados dos usuários, e armazenar no banco de dados. Sempre que alguém quiser aprender um fato estatístico sobre a população em questão, deve fazer uma pergunta, ou *query*, ao curador. Assim, é seu papel receber a pergunta e calcular a resposta da *query*. Por fim, para satisfazer a privacidade diferencial, é necessário perturbar os dados com uma função de randomização, adicionando ruído [10].

Nas *queries* de contagem (*count*), o usuário define o parâmetro que será usado para decidir se um registro faz parte dessa contagem. O curador então percorre todos os registros, aumentando em 1 a contagem quando encontra um registro que atende o parâmetro passado. Ao final, um ruído é aplicado ao resultado de forma a obter-se uma resposta diferencialmente privada antes de retorná-la ao usuário [5]. Como exemplo, em um banco de dados onde os registros possuem informação de localização, podemos usar

como parâmetro para a contagem o país onde aquela localização está. Dessa forma, ao responder a *query*, teremos quantos registros são de cada país.

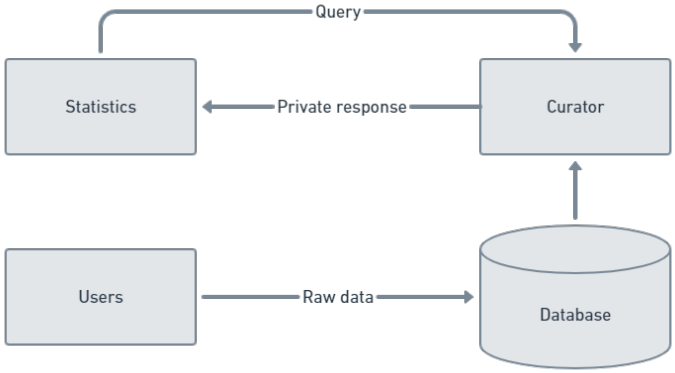


Figura 1 – Representação do funcionamento da privacidade diferencial central. (Fonte: [1, p. 2])

Como é possível observar na Figura 1, o curador possui acesso ao banco de dados com as informações reais dos usuários, e as utiliza para gerar um resultado com privacidade diferenciada, que responde as *queries* necessárias para gerar estatísticas.

A função de randomização tem como objetivo perturbar um dado, aplicando ruído nele. Para isso, normalmente, usa-se da aleatoriedade para decidir se o dado será alterado e qual é o valor resultante [5]. Como exemplo, uma função que gera números aleatórios entre 1 e 10, pode definir um escopo e, caso o número gerado esteja dentro dele, o valor do resultado será alterado.

Porém, além da aleatoriedade, é preciso criar uma probabilidade maior de que o dado não seja alterado, de forma que o resultado ainda seja útil. No exemplo anterior, a probabilidade do dado continuar o mesmo foi de $\frac{7}{10}$, o que permite que o valor passado não será alterado na maior parte das vezes.

Como mencionado anteriormente, para atingir a privacidade diferencial, o curador precisa aplicar uma função de randomização κ, tendo como entrada um banco de dados e como saída a informação diferencialmente privada. Dizemos então que uma função κ satisfaz $\epsilon - DP$ se para os bancos de dados D_1 e D_2 diferindo em apenas um registro, e onde S são todos os resultados possíveis da função, a inequação (2.1) é verdadeira [10].

$$Pr[k(D_1) \in S] \leq exp(\epsilon) \times Pr[k(D_2) \in S] \tag{2.1}$$

O objetivo da inequação é garantir que a presença ou ausência de um único registro em um banco de dados não altere o resultado da função κ, de forma que não seja possível inferir a presença deste registro.

O parâmetro ϵ, chamado *privacy budget*, ou *privacy loss* (perda de privacidade), permite definir a "quantidade de privacidade" que a definição impõe. À medida que ϵ diminui, os resultados obtidos de κ para entradas similares devem ficar mais parecidos, dificultando ainda mais encontrar a diferença entre eles e aumentando o nível de privacidade. De forma inversa, conforme ϵ aumenta, os resultados podem ser mais diferentes um do outro, diminuindo o nível de privacidade.

Para satisfazer $\epsilon - DP$, conforme sua inequação, as funções de randomização utilizam o parâmetro ϵ para definir quanto ruído será aplicado nos dados. Quanto maior o ϵ, menos ruído será aplicado, e o resultado ficará mais parecido com os valores de entrada.

O uso de uma mesma função de randomização sobre um mesmo banco de dados, repetidas vezes, pode permitir que se saiba o valor verdadeiro do resultado, pois a média das respostas tende a convergir ao valor real. Porém, ainda é garantido a impossibilidade de inferência da presença de um registro [5].

2.2 Privacidade Diferencial Local

Como visto na seção anterior, a privacidade diferencial tradicional, também chamada de privacidade diferencial central, baseia-se em um coletor e um curador confiáveis que têm acesso às informações verdadeiras do usuário. No entanto, muitas vezes, um usuário não se sente confortável em confiar suas informações a terceiros, uma vez que o banco de dados com os dados reais ainda existe e mesmo grandes empresas já tiveram os dados de usuários expostos [11].

Para resolver esse problema, foi definida a privacidade diferencial local [12], uma configuração de DP que garante ainda mais controle dos usuários sobre os próprios dados. Nesse modelo, cada usuário possui e mantém suas próprias informações reais. Porém, ao invés de enviar os dados originais ao coletor, eles passarão por uma função randomizadora antes e então poderão ser repassados, já perturbados, ao coletor de dados. Dessa forma, não há necessidade de confiar em terceiros, uma vez que apenas o usuário terá acesso à resposta original sem ser perturbada pelo algoritmo. O banco de dados do coletor agora possui apenas as informações diferencialmente privadas dos usuários, como apresentado na Figura 2 e, dessa forma, mesmo que ocorra um vazamento do banco, o usuário não terá seus dados expostos.

Como neste cenário não temos os dados reais em um banco, não há mais a necessidade da figura do curador de dados confiável. Isso permite que as *queries* sejam feitas diretamente ao banco. No entanto, diferente do modo central, o ruído foi aplicado levando em conta apenas um usuário, o que pode dificultar a interpretação do resultado da contagem.

Para diminuir o ruído aplicado no dispositivo de cada usuário, sempre que uma

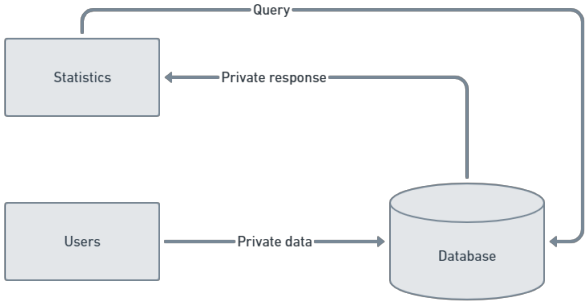


Figura 2 – Representação do funcionamento da privacidade diferencial local. (Fonte: [1, p. 2])

pergunta é feita, um agregador aplicará a função agregadora. Essa função leva em conta o número de respostas e o resultado da *query*, tratando os dados como um conjunto, gerando uma estimativa do resultado e garantindo a imparcialidade estatística [1].

Para responder uma *query* com LDP, é necessário que três processos sejam aplicados [13]. Primeiro, os dados reais de cada usuário são codificados em um novo formato, como um vetor de bits, por exemplo. Em seguida, o dado codificado é perturbado pelo algoritmo randomizador, tornando-se diferencialmente privado, e pronto para envio ao coletor. Após o coletor receber os dados perturbados de cada usuário, cabe ao agregador aplicar uma função agregadora e responder a *query*.

Para uma função satisfazer $\epsilon - LDP$, ela deve seguir também a inequação (2.1), com a diferença que no caso da LDP, os bancos de dados contém apenas um registro, pois será aplicado por cada usuário sem informações dos outros, diferenciando do modelo central, cuja inequação se baseia em um banco de dados de n registros. Normalmente, as funções se baseiam em um método chamado *Randomized Response*, que aposta na aleatoriedade enviesada como forma de responder algo sensível, sem garantir que respondeu-se a verdade, porém dando mais chance da resposta ser a verdadeira [14].

Como exemplo, se uma pessoa quer responder uma pergunta de “Sim” ou “Não” sem que saibam se a resposta é verdadeira e, mesmo assim, garantir o valor estatístico da pesquisa, sugere-se o seguinte: antes de responder, o respondente lança uma moeda. Caso o resultado seja “cara”, o respondente deve dar a resposta correta. No entanto, se o resultado for “coroa”, ele deve jogar novamente a moeda. Nesse segundo arremesso, a resposta verdadeira é ignorada e, caso o resultado seja “cara” deverá responder “Sim”, e se o resultado for “coroa” deverá responder “Não”.

Para o caso em que o indivíduo responder “Sim”, é impossível saber se esta resposta foi dada por ser a verdade ou por ter obtido um resultado “coroa” no primeiro lançamento e “cara” no segundo. Além disso, no primeiro arremesso, a probabilidade de cair “cara” e

a resposta ser verdadeira é $\frac{1}{2}$. Se cair "coroa", a chance de cair "cara" e a resposta "Sim" ser a verdadeira é de $\frac{1}{4}$. Dessa forma, a chance da resposta ser verdadeira é $\frac{3}{4}$, garantindo que ao agregar várias respostas será possível aprender algo real sobre os dados, sem saber quem respondeu a verdade ou quem mentiu [5].

Obviamente, esta técnica funciona apenas para casos com duas opções de resposta, como perguntas de "Sim" ou "Não" [13] [1]. Em qualquer caso da vida real onde a possibilidade de resposta for maior que duas, precisamos de um algoritmo mais complexo. Existem muitos mecanismos que utilizam como base esse conceito de resposta randomizada, e que são expandidos para conseguir várias possíveis respostas. Como exemplo, temos o RAPPOR, desenvolvido pela Google, que usa o conceito de resposta randomizada generalizada [15].

3 MATERIAIS E MÉTODOS

O objetivo deste trabalho é avaliar a privacidade diferencial em um cenário onde temos dados de usuários que alugam bicicletas através de um programa de bicicletas compartilhadas promovido pelo site "Citi Bike", na cidade de Nova Iorque, nos Estados Unidos. Mais especificamente, busca-se observar a relação entre privacidade e utilidade dos dados quando são aplicadas técnicas de DP central e LDP.

As técnicas aplicadas tiveram o objetivo de responder *queries* de contagem (*count*), onde o resultado retorna quantos elementos do banco de dados possuem uma propriedade P.

O serviço de bicicletas coleta informações de todos os aluguéis e disponibiliza o conjunto de dados[1] com essas informações ao público. Dentre os dados disponíveis, é possível encontrar informações como as estações de onde a bicicleta saiu e chegou e suas coordenadas geográficas, além do horário do início e fim da corrida. As informações também contêm gênero e ano de nascimento do usuário, e até mesmo se ele é membro do serviço ou não.

Percebe-se que não há nenhum dado que realmente identifique o usuário, como um nome, idade ou endereço de residência. No entanto, como visto previamente, é possível combinar as informações disponíveis, como gênero e data de nascimento de uma pessoa, com dados adicionais ou informações anteriores, permitindo a identificação de quem fez o passeio. Tendo em vista essas possibilidades, verifica-se a importância de aplicar técnicas de privacidade diferencial para buscar garantir a privacidade dos usuários do serviço.

3.1 Métodos de privacidade diferencial utilizados

Como explicado anteriormente, a privacidade diferencial possui dois modos. No modo central, existe um curador que possui acesso aos dados reais, e aplica uma função de randomização, que satisfaz privacidade diferencial, para responder *queries*. Já no modo local, cada usuário tem a função que satisfaz a privacidade diferencial aplicada nos seus dados diretamente de seu dispositivo, ou seja, a figura do curador só teria acesso aos dados já manipulados. Essa diferença garante uma maior proteção para os sistemas com LDP, pois mesmo que o banco de dados do coletor seja acessado ilegalmente, não é possível identificar indivíduos e seus dados neste banco.

Assim, é importante comparar os dois métodos, de forma a entender as vantagens e desvantagens que essa maior proteção da LDP tem em relação a DP central.

[1] <https://ride.citibikenyc.com/system-data>

3.1.1 Privacidade Diferencial Central

Para realizar os experimentos no modo central, foi utilizada a implementação de uma das formas mais comuns de funções para atingir a privacidade diferencial, o mecanismo de *Laplace* [5]. Nesse mecanismo, o ruído gerado é definido pela distribuição de *Laplace*, que é calculada através da Equação 3.1, com escala b, e obtém sua incerteza através de uma variável x gerada aleatoriamente.

$$Lap(x|b) = \frac{1}{2b} exp(-\frac{|x|}{b}) \tag{3.1}$$

Assim, para um conjunto de dados D, com uma função $f(D)$, que possui sensitividade s e retorna um número, a seguinte definição do mecanismo de *Laplace* satisfaz $\epsilon - DP$:

$$F(D) = f(D) + Lap(\frac{s}{\epsilon}) \tag{3.2}$$

Onde *Lap* é uma função que retorna um valor aleatório da distribuição de *Laplace*, calculada com média $\frac{s}{\epsilon}$.

A sensitividade de uma função representa o quanto o resultado daquela função muda quando a entrada é alterada em "um". Por exemplo, em uma função que realiza a contagem de indivíduos em um banco de dados, a sensitividade é 1, pois ao adicionar, ou remover, um indivíduo no banco, alterando a entrada em um registro, a contagem só pode aumentar ou diminuir em 1. Como todos os experimentos utilizaram a *query* de contagem, a sensitividade se manteve constante em 1.

Tratando-se do modo central, é necessário que os dados estejam disponíveis para o curador. Na implementação feita, os registros do aluguel de bicicletas foram salvos em uma coleção de objetos. Para responder as *queries* de contagem feitas, primeiro é decidido o parâmetro que vai definir se um usuário pertence ou não ao resultado de uma contagem. Com o parâmetro definido, os dados são acessados em sequência para encontrar qual a contagem em que aquele objeto será adicionado. Quando encontrada, seu valor é aumentado em 1.

Após o cálculo de todas as contagens, é criada uma coleção com os valores resultantes. Para cada um desses valores, será aplicado o mecanismo de *Laplace*, que irá garantir que aquele valor seja diferencialmente privado. Por fim, a coleção resultante da aplicação de DP nos valores das contagens é usada para escrever uma planilha que será manejada para gerar os gráficos dos experimentos, onde cada linha possui o identificador da contagem e seu valor.

3.1.2 Privacidade Diferencial Local

Diferente do modo central, a LDP perturba os dados de cada usuário em seu próprio dispositivo. Teoricamente, os mecanismos utilizados para alcançar a DP podem ser usados por cada usuário antes de enviar os dados para o servidor. Porém, em sua maioria, o funcionamento só acontece com o acesso a todos os registros do banco, impedindo o uso local [16].

Apesar da resposta randomizada já satisfazer $\epsilon - LDP$, ela só permite respostas de "Sim" ou "Não". Desta forma, para realizar os experimentos no modo local, foi feita uma implementação baseada em *Programming Differential Privacy* [8], que fornece um algoritmo que utiliza *unary encoding*, seguindo a abordagem feita por Wang et al. [17].

O *unary encoding* é um mecanismo que usa a codificação dos valores em um vetor de zeros e uns. Esse vetor é perturbado e, então, pode ser enviado para o coletor de dados e ser agregado para calcular as estatísticas. Para que os três processos, codificar, perturbar e agregar, funcionem corretamente, primeiro, é preciso definir o domínio das respostas, ou seja, quais as respostas possíveis para a pergunta a ser feita. Para um domínio de tamanho n, cada usuário terá suas informações codificadas em um vetor de n posições, onde cada posição representa uma resposta possível. Apenas a posição referente à resposta do usuário terá o valor 1, e todas as outras terão o valor 0.

O próximo passo, no qual perturba-se o vetor, é feito invertendo o valor de algumas posições. Para escolher quais posições terão o valor invertido, é gerado um número aleatório para cada uma delas. Caso o valor da posição seja 1 e o número gerado seja menor ou igual a p, então aquela posição manterá o valor 1 no vetor resultante. Para o caso do valor ser 0, o número gerado deverá ser menor ou igual a q para que a posição obtenha o valor 1. A probabilidade de uma posição ter o valor 1 é obtida pela Equação 3.3, onde B é o vetor codificado e B' é o vetor resultante após a perturbação.

$$\mathsf{Pr}[B'[i] = 1] = \begin{cases} p & se\ B[i] = 1 \\ q & se\ B[i] = 0 \end{cases} \tag{3.3}$$

O uso do número aleatório em casos que p é maior que q gera incerteza sobre o resultado, mas garante que haja maior probabilidade do valor resultante ser verdadeiro e que o protocolo satisfaça $\epsilon - LDP$, o qual é obtido pela Equação 3.4.

$$\epsilon = \log\left(\frac{p(1-q)}{(1-p)q}\right) \tag{3.4}$$

Os dois primeiros passos, codificar e perturbar, devem ser feitos localmente no dispositivo do usuário. No entanto, nos experimentos realizados, eles foram aplicados em

cada registro do banco de dados utilizado, simulando usuários que tinham suas informações perturbadas para enviar para o coletor.

Após codificar e perturbar todos os registros, os vetores são salvos em um outro vetor, representando o resultado obtido ao coletar os dados já protegidos de todos os usuários. As contagens são obtidas somando o valor de cada posição de todos os vetores, resultando, assim, em um vetor com a contagem total de cada resposta em sua devida posição. Porém, como a perturbação dos dados faz com que obtenhamos valores errados na contagem, é aplicado um último passo.

A agregação leva em conta o número de respostas obtidas e a soma delas, ou seja, a contagem obtida. Assim, o vetor final com as contagens de cada resposta é obtido pela Equação 3.5, onde B'_j é o vetor perturbado de cada usuário, ou seja, cada posição do vetor resultante é influenciada pela soma dos valores daquela posição em todas as respostas perturbadas e pelo número de respostas, além de p e q.

$$A[i] = \frac{\sum_j B'_j[i] - nq}{p - q} \tag{3.5}$$

Por fim, esse vetor com as contagens agregadas é usado para escrever uma planilha, onde cada linha possui um identificador da resposta e sua contagem, que, assim como no método central, será utilizada para gerar os gráficos dos experimentos.

3.2 Experimentos

Em todos os experimentos, serão executados dois programas, um utilizando da privacidade diferencial central, e o outro utilizando da local. Cada um dos programas será executado 3 vezes, diferindo neles apenas o ϵ escolhido. Um deles será feito com o valor 0,008, um número muito baixo e que garante alta privacidade, porém resultados mais imprecisos. Outra execução será feita com o valor 2,251, sendo um valor médio. E por fim, será usado o valor 13,814, representando um valor alto que diminui bastante a probabilidade de que um dado tenha seu valor alterado, garantindo um resultado mais parecido ao dos dados reais. Esses valores foram obtidos através da Equação 3.4, com p igual a 0,501, 0,755 e 0,999 respectivamente.

O objetivo dos programas é responder *queries* do tipo *count*, como "Quantos aluguéis de bicicleta foram feitos em cada hora do dia?". Como resultado, serão geradas planilhas com a resposta da *query*, contendo identificadores da contagem e seu valor.

Quando aplicamos a privacidade diferencial, o resultado das consultas como o *count* é uma aproximação do que seria a mesma consulta realizada sobre os dados reais. Como toda aproximação, ela acaba apresentando um erro com relação ao dado real. Para medirmos este erro, podemos usar diferentes métricas, baseadas no erro absoluto ou no

erro percentual. A seguir, damos mais detalhes sobre as métricas de erro que utilizamos em nosso trabalho.

Para o primeiro experimento, foi usado o erro absoluto médio (MAE - *mean absolute error*), cuja fórmula é (3.6). Nela, temos y_i que representa o valor real, e $\hat{y}_i$ que representa o valor aproximado, além da quantidade de valores n. Seu objetivo é encontrar o erro absoluto, que é calculado pela diferença entre o valor real e o estimado, para todos os valores, e então, somá-los e dividir pela quantidade de valores, encontrando a média dos erros absolutos.

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.6}$$

Para o segundo e terceiro experimentos, houve a necessidade de outro tipo de cálculo para facilitar a comparação de resultados. Para isso, foi utilizado o erro percentual. Sua fórmula (3.7) utiliza o erro absoluto e então divide pelo valor verdadeiro, e após multiplicar por 100 resulta em um percentual.

$$100\%\frac{|y - \hat{y}|}{y} \tag{3.7}$$

No entanto, para casos onde o número real é 0, esse cálculo cria um problema de divisão por zero. Isso ocorre principalmente nos experimentos onde foi feita a contagem por hora e por gênero. Para resolver esse problema, utilizamos uma fórmula ajustada (3.8), onde o percentual é obtido pela média do valor esperado com o valor real. Ainda existe um problema quando ambos os valores são zero, porém, nesses casos, o erro retornado é 0%, pois os valores são iguais. Por fim, é importante ressaltar que o valor varia entre 0% e 200%.

$$100\%\frac{|y - \hat{y}|}{(|y| + |\hat{y}|)/2} \tag{3.8}$$

Para aplicar os cálculos e obter as estatísticas, foi utilizada a biblioteca Scikit Learn [18], voltada para a linguagem Python. Dela, foi usada a função para o MAE, e foi feita uma adaptação do código do erro relativo para conseguir o ajustado, conforme a fórmula (3.8).

4 RESULTADOS

4.1 Experimento 1: Erro absoluto médio

Neste experimento, a contagem realizada é a de quantas mulheres alugaram uma bicicleta em cada estação durante cada hora do dia. Depois, é calculado o erro absoluto médio dos resultados com privacidade diferencial, em comparação com o resultado real. Para chegar no resultado dessa contagem foram utilizados apenas dois parâmetros, as estações e as horas, de forma que foi possível obter a contagem de usuários que alugaram bicicletas em cada estação durante as horas do dia. Para conseguir filtrar ainda mais as possibilidades de usuários, apenas registros de usuários do gênero feminino foram usados.

Para esse experimento, utilizamos um banco de dados com 41 mil registros, o qual contém dados suficientes para todas as horas do dia.

Como explicado anteriormente, o experimento foi executado seis vezes com diferentes parâmetros todas as vezes. Obteve-se, ao fim de cada execução, as contagens privadas, podendo então comparar com os dados reais e calcular o MAE de cada execução.

Tabela 1 – Erro absoluto médio.

ϵ	**Método**	**MAE**
0,008	Central	63,74701
0,008	Local	11259,62888
2,251	Central	0,35796
2,251	Local	37,10306
13,814	Central	0,00119
13,814	Local	2,29459

Como podemos ver na Tabela 1, tanto o ϵ quanto o método influenciam muito no erro absoluto médio. Conforme ϵ aumenta, o MAE diminui, mostrando como os dados protegidos se tornam mais próximos ao original e, consequentemente, menos privados.

Também é possível entender a diferença entre os dois métodos, central e local, onde com $\epsilon = 0,008$, o MAE no método local foi muito alto, principalmente se comparado com o do central. Esse resultado mostra como apesar do método local oferecer menos riscos ao dar apenas os dados já privados ao coletor, ele necessita de um ϵ maior se quiser comparar o nível de utilidade dos seus dados ao método central.

Além disso, pelo fato da *query* pedir a contagem de aluguéis feitos por mulheres em cada estação por hora, a contagem de cada possibilidade gerou resultados com números baixos, na casa das dezenas, dificultando ainda mais para o método local, que necessita de dados de vários indivíduos para melhorar a sua utilidade.

4.2 Experimento 2: Erro percentual durante as horas do dia

Neste experimento, é realizada a contagem do número de pessoas que alugaram bicicletas em cada hora do dia. Depois, é calculado o erro percentual dos resultados com privacidade diferencial, em comparação com o resultado real. A contagem foi feita utilizando apenas a hora em que o aluguel foi feito, independente do dia ou das outras informações do usuário. Assim conseguimos verificar as horas do dia em que o serviço é mais utilizado, e como essa frequência influencia na utilidade das informações geradas. Para o experimento, o banco de dados utilizado contém 100 mil registros de aluguéis.

Após as 6 execuções, foram obtidas as contagens em cada horário do dia, de cada método e ϵ. Por fim, o erro percentual do resultado das contagens é calculado e apresentado em gráficos, onde é possível observar a imprecisão dos dois métodos ao longo do dia para cada ϵ.

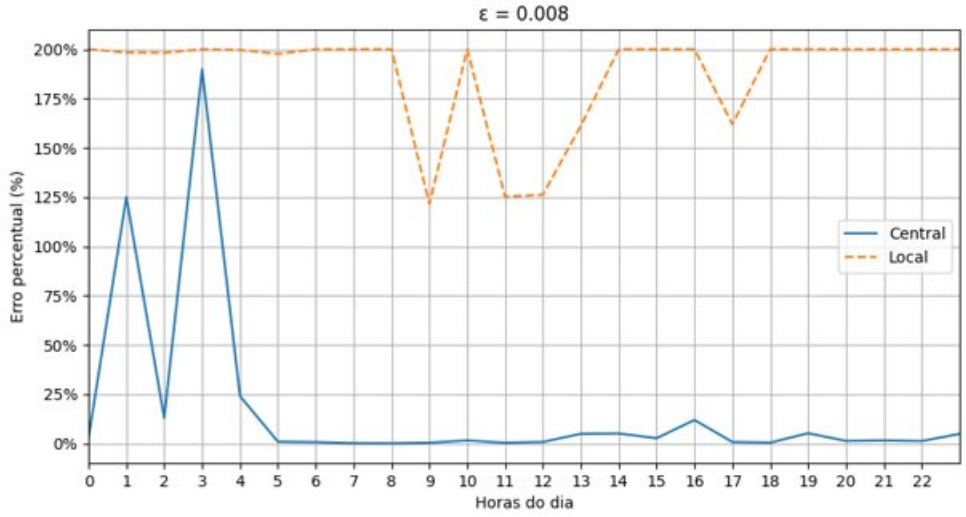


Figura 3 – Erro percentual calculado em cada hora do dia com $\epsilon = 0{,}008$

Ao analisar os resultados da contagem não privada, foi possível observar que nas horas iniciais do dia, das 0h às 05h, além das últimas horas do dia, o uso do serviço é muito baixo se comparado ao resto do dia. Esse baixo número de usuários causa um efeito percebido no gráfico da Figura 3, onde o erro percentual é maior durante os horários citados.

Além disso, em comparação com o experimento 1, nesses momentos onde há poucos registros, os métodos central e local possuem erros mais parecidos, também devido ao percentual máximo de 200%, mas que mostra que apesar de eficaz, o algoritmo de DP central, ao usar um ϵ baixo, também precisa de um alto número de indivíduos para que os dados protegidos tenham maior utilidade.

No caso do LDP, apenas nos horários de pico dos registros foi possível perceber maior utilidade nos dados com a queda do erro percentual, mas em geral a utilidade dos dados nesse nível de privacidade continua muito baixa.

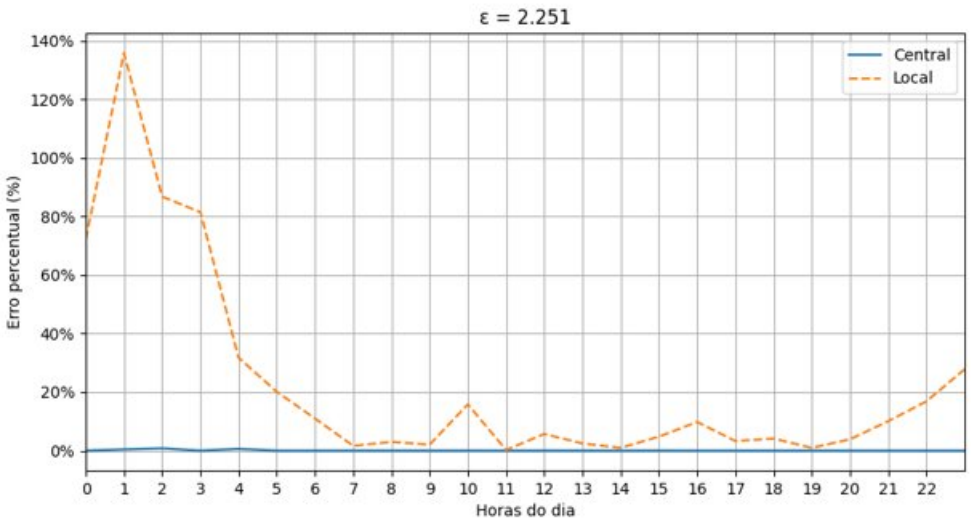


Figura 4 – Erro percentual calculado em cada hora do dia com $\epsilon = 2{,}251$

Já no gráfico da Figura 4, com $\epsilon = 2{,}251$, é possível ver como o método central consegue alcançar um valor do erro percentual próximo de 0 em todas as horas do dia, o que se mantém no gráfico da Figura 5, com $\epsilon = 13{,}814$. Essas informações também corroboram os resultados do primeiro experimento, onde considerando ϵ médio, o MAE chega perto de 0.

Além disso, no caso do método local, o erro também caiu, chegando a um máximo de 9% na Figura 5. No entanto, a discrepância entre os períodos de menor uso do serviço em comparação com as outras horas do dia continuou. Já com $\epsilon = 2{,}251$, o erro percentual do método local nos horários de pico não passou de 20%, chegando próximo de 0 e mostrando conseguir chegar bem próximo da utilidade do método central, quando possui um alto número de indivíduos na base de dados.

4.3 Experimento 3: Erro percentual durante as horas do dia nas estações mais frequentadas

Neste experimento, a contagem realizada é da quantidade de aluguéis de bicicleta em cada estação durante cada hora do dia. Como é semelhante ao primeiro experimento, os processos aplicados para obter os resultados foram os mesmos, exceto pela filtragem por gênero, usando todos os usuários disponíveis. Além disso, já foi possível observar fatos importantes sobre a diferença do resultado entre os modos de privacidade diferencial, e

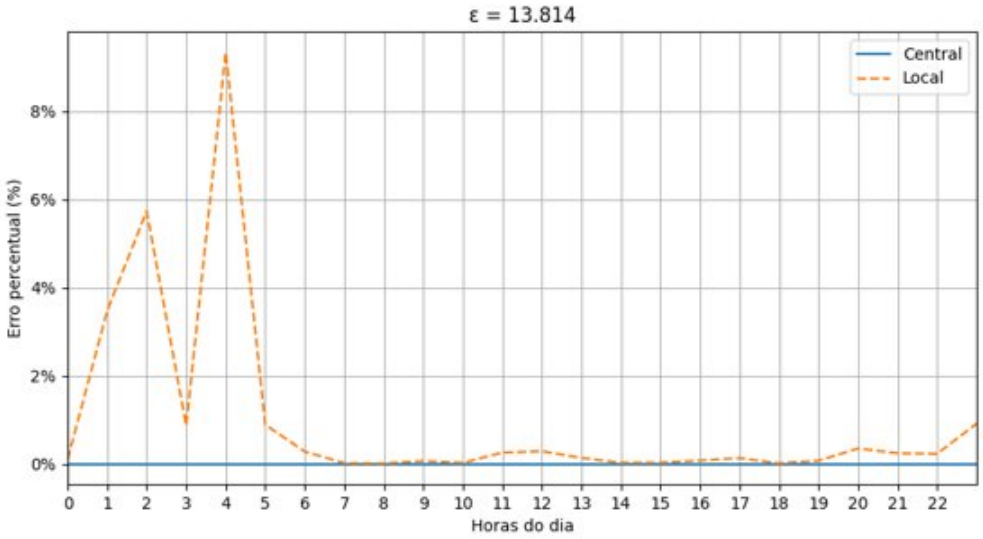


Figura 5 – Erro percentual calculado em cada hora do dia com $\epsilon = 13.814$

sobre como reagem a cada ϵ, entendendo como o número de registros e o método influenciam na utilidade e precisão do resultado. Assim, foi utilizado para a discussão apenas o resultado do erro percentual da contagem no modo central com ϵ médio.

Após a execução, as 24 estações com mais aluguéis registrados foram separadas em outra planilha. Depois, foi realizado o cálculo do erro percentual de cada resultado com privacidade diferencial em comparação com os dados reais. O resultado desse cálculo foi utilizado para gerar o mapa de calor apresentado na Figura 6.

Nele, é possível observar que mesmo com um ϵ médio, algumas estações, em determinadas horas, possuem o erro percentual alto, com valores acima de 100%. Mesmo nas estações mais frequentadas, conforme cai o movimento, mais erros são identificados, principalmente na madrugada. No entanto, a média de erro das estações ao longo do dia diminui conforme o número total de aluguéis aumenta.

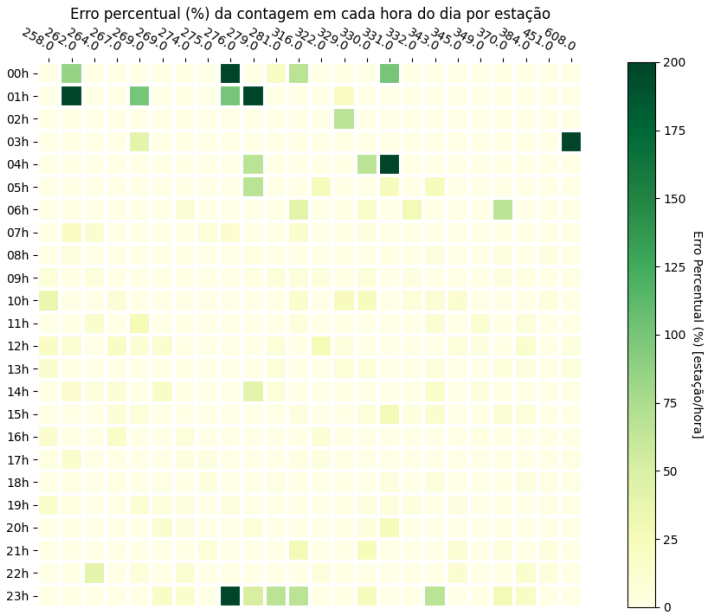


Figura 6 – Mapa de calor com o erro percentual obtido no experimento 3

5 CONCLUSÃO

A privacidade torna-se cada vez mais importante, e por isso, é amplamente debatida e divulgada, gerando diversas discussões sobre como preservá-la. Neste trabalho, sugere-se a privacidade diferencial como uma forma de evitar que usuários de serviços digitais sejam prejudicados por disponibilizar seus dados a serviços online.

Além disso, foram propostas duas aplicações da DP, sendo a primeira no método central, onde o serviço coleta dados reais dos usuários e aplica ruído nesses dados para responder perguntas estatísticas, como contagem dos dados. A segunda aplicação utiliza de privacidade diferencial local, que diferentemente do central, consiste na coleta dos dados dos usuários já com ruídos, de forma que apenas o usuário tenha acesso aos dados reais, garantindo que mesmo com a invasão do banco de dados do serviço, os dados já estarão diferencialmente privados.

Os experimentos deste trabalho foram realizados através dos registros de aluguel de bicicletas do site "Citi Bike", que promove um programa de bicicletas compartilhadas na cidade de Nova Iorque, Estados Unidos. Por meio deste, foram obtidas informações relacionadas aos horários em que os usuários mais utilizam o serviço, ao movimento das esta ções ao longo do dia e ao número de alugueis feitos por usuários do gênero feminino ao longo do dia.

Essas informações coletadas podem ser usadas para ferir a privacidade dos usuários através de vazamento de dados, o que acarretaria em perdas de usuários para o site e também aplicação de ações judiciais. Por conseguinte, considerando os prejuízos que a falta de privacidade dos dados pode trazer para o serviço e para os usuários, mostra-se a importância da utilização da privacidade diferencial para proteger essas informações.

Adicionalmente, os experimentos permitiram observar o comportamento dos dois modos de privacidade diferencial, ao tentar realizar consultas de contagem nos dados de aluguel. Apesar de, no modo central, o usuário precisar disponibilizar seus dados reais ao serviço, o resultado obtido é muito mais próximo do real que o gerado pelo modo local, o que demonstra que apesar de garantir a segurança ao usuário de que apenas ele sabe a informação real, é necessária uma perda de privacidade (ϵ) muito maior para ser tão útil quanto o modo central.

Além disso, em consultas onde o número de registros é pequeno, o erro das respostas geradas por ambos os modos aumenta, demonstrando como aplicações que usam privacidade diferencial conseguem resultados mais úteis conforme aumenta o número de indivíduos no conjunto de dados.

Em suma, todos os experimentos deste trabalho utilizaram a contagem de registros

para obter seus resultados. Para trabalhos futuros, outros tipos de estatísticas poderiam ser obtidas, como médias, com o objetivo de entender o comportamento dos métodos em diferentes cenários.

REFERÊNCIAS

[1] YANG, M. et al. Local differential privacy and its applications: A comprehensive survey. *arXiv preprint arXiv:2008.03686*, 2020.

[2] FRIEDMAN, A.; SCHUSTER, A. Data mining with differential privacy. In: *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.: s.n.], 2010. p. 493–502.

[3] ZHAO, X. et al. Ldpart: Effective location-record data publication via local differential privacy. *IEEE Access*, v. 7, p. 31435–31445, 2019.

[4] LIU, L. From data privacy to location privacy: Models and algorithms. In: CITESEER. *VLDB*. [S.l.], 2007. v. 7, p. 1429–1430.

[5] DWORK, C.; ROTH, A. et al. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, v. 9, n. 3-4, p. 211–407, 2014.

[6] BÉLANGER, F.; CROSSLER, R. E. Privacy in the digital age: a review of information privacy research in information systems. *MIS quarterly*, JSTOR, p. 1017–1041, 2011.

[7] FAINMESSER, I. P.; GALEOTTI, A.; MOMOT, R. Digital privacy. *HEC Paris Research Paper No. MOSI-2019-1351*, 2019.

[8] NEAR, J. P.; ABUAH, C. *Programming Differential Privacy*. [s.n.], 2021. v. 1. Disponível em: <https://uvm-plaid.github.io/programming-dp/>.

[9] DWORK, C. Differential privacy. In: SPRINGER. *International Colloquium on Automata, Languages, and Programming*. [S.l.], 2006. p. 1–12.

[10] DWORK, C. Differential privacy: A survey of results. In: SPRINGER. *International conference on theory and applications of models of computation*. [S.l.], 2008. p. 1–19.

[11] ZUO, C.; LIN, Z.; ZHANG, Y. Why does your data leak? uncovering the data leakage in cloud from mobile apps. In: IEEE. *2019 IEEE Symposium on Security and Privacy (SP)*. [S.l.], 2019. p. 1296–1310.

[12] KASIVISWANATHAN, S. P. et al. What can we learn privately? *SIAM Journal on Computing*, SIAM, v. 40, n. 3, p. 793–826, 2011.

[13] WANG, T. et al. Locally differentially private protocols for frequency estimation. In: *26th {USENIX} Security Symposium ({USENIX} Security 17)*. [S.l.: s.n.], 2017. p. 729–745.

[14] WARNER, S. L. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, Taylor & Francis, v. 60, n. 309, p. 63–69, 1965.

[15] ERLINGSSON, Ú.; PIHUR, V.; KOROLOVA, A. Rappor: Randomized aggregatable privacy-preserving ordinal response. In: *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*. [S.l.: s.n.], 2014. p. 1054–1067.

[16] QIN, Z. et al. Heavy hitter estimation over set-valued data with local differential privacy. In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. [S.l.: s.n.], 2016. p. 192–203.

[17] WANG, T. et al. Optimizing locally differentially private protocols. *arXiv preprint arXiv:1705.04421*, 2017.

[18] BUITINCK, L. et al. API design for machine learning software: experiences from the scikit-learn project. In: *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*. [S.l.: s.n.], 2013. p. 108–122.