Capítulo 3

Preservação de Privacidade de Dados: Fundamentos, Técnicas e Aplicações

Felipe T. Brito, Javam C. Machado

Abstract

Many organizations have perform data analysis over published data to find hidden patterns and foresee future tendencies. The data mining process is only possible because data can be reached through query services or open access to published data. However, published data may carry out information that uniquelly identify individuals, which may lead to privacy violation. Keeping the utility of the data for mining and still maintaining individual's privacy is a scientific problem that has been studied over the past few years. The objective of this chapter is to present the main concepts of data privacy preserving and to describe the techniques capable of assuring that no one can be reidentified from their publish data. Additionally it describes real world applications that apply these techniques as a way of preserving individual's privacy, while keeping data utility for further data analysis when requested.

Resumo

Muitas organizações realizam análises importantes sobre dados a fim de descobrir padrões ocultos e prever tendências futuras. Para que muitas dessas análises sejam realizadas, é necessário que os dados estejam disponíveis para acesso, seja por meio de publicações ou de serviços de consulta. Entretanto, dados acessíveis pelo público podem conter informações que identificam unicamente indivíduos, causando assim uma violação de privacidade. Manter a utilidade dos dados para que análises sejam realizadas e, simultaneamente, garantir a privacidade dos indivíduos é um problema que tem recebido bastante atenção nos últimos anos. Este capítulo tem por objetivo apresentar os principais conceitos em torno da preservação de privacidade de dados, além das técnicas para assegurar que indivíduos não possam ser reidentificados a partir do compartilhamento de suas informações. Adicionalmente, são demonstradas aplicações em cenários reais que utilizam as técnicas apresentadas como forma de preservação de privacidade, enquanto buscam reter a maior quantidade de informação possível para eventuais análises.

3.1. Introdução

Atualmente, grandes volumes de dados têm sido coletados por governos, corporações ou mesmo instituições ao redor do mundo. Dados são bens muito valiosos para as organizações dos mais diversos tipos, sejam elas bancárias, de seguros, de varejo, de saúde, etc. Por exemplo, uma base de dados de serviços de saúde pode conter informações sobre as reações de seus pacientes a um determinado medicamento ou tratamento. Essas informações podem ser úteis para companhias farmacêuticas fabricarem e distribuírem medicamentos. Bancos também gostariam de entender como seus clientes estão utilizando cartões de crédito, para assim oferecerem outros tipos de serviços com confiança de crédito. Seria ideal que comércios e supermercados pudessem entender o comportamento de seus clientes, e assim ofertarem produtos e serviços de maneira mais assertiva. Outro exemplo seria um conjunto de dados coletado por iniciativas privadas que contém informações sobre mobilidade urbana. Essas informações podem ser muito bem aproveitadas para identificar e prever concentrações de fluxos de automóveis, melhorar o transporte público, planejar obras de infraestrutura, entre outros benefícios. Muitos desses padrões estão escondidos na grande quantidade de dados coletados.

A mineração de dados é o processo de extração de informações, não conhecidas a priori, a partir de grandes conjuntos de dados [28]. O sucesso da mineração ocorre devido à disponibilidade dos dados com qualidade e ao compartilhamento efetivo das informações. Dessa forma, é possível realizar análises para descobrir padrões ocultos e/ou para prever tendências futuras, permitindo tomadas de decisão baseadas no conhecimento e não apenas em opiniões próprias ou intenções. Muitas informações coletadas podem servir para que empresas forneçam serviços de valor agregado para seus clientes, o que, por sua vez, resulta em maior receita e consequentemente em maior desenvolvimento para a sociedade. Para realizar um processo de análise eficiente é necessário que os dados estejam disponíveis de alguma forma. Se não há dados confiáveis em quantidade suficiente para a realização de pesquisas e análises, torna-se obscuro o uso da informação em prol do desenvolvimento da sociedade.

Uma forma de disponibilização e utilização de dados que tem ganhado bastante visibilidade nos últimos anos é o modelo de dados abertos [27]. Nesse modelo, dados são geralmente disponibilizados por instituições governamentais, em um formato computacional aberto e processável por máquina, isto é, que não possui restrição tecnológica para ser utilizado, a fim de permitir cidadãos comuns, empresas, instituições de ensino e organizações não-governamentais utilizá-los de maneira inovadora, gerando valor para a sociedade. Dessa forma, dados abertos são considerados fundamentais para acelerar o progresso da ciência e possibilitar investimentos em educação, saúde, segurança pública, transporte, uso racional de recursos e desenvolvimento sustentável. Dentre os benefícios dessa prática destacam-se: maior transparência sobre as atividades governamentais; acompanhamento das políticas públicas; melhoria da troca de informações entre órgãos e esferas de governo; incentivo à sociedade em desenvolver soluções para o bem comum; fomento a novos mercados de tecnologia da informação; estímulo a inovação e pesquisa; entre muitos outros.

Sob outra perspectiva, ao mesmo tempo que a disponibilização de dados para análises e descoberta de padrões é fundamental para o desenvolvimento da sociedade contem-

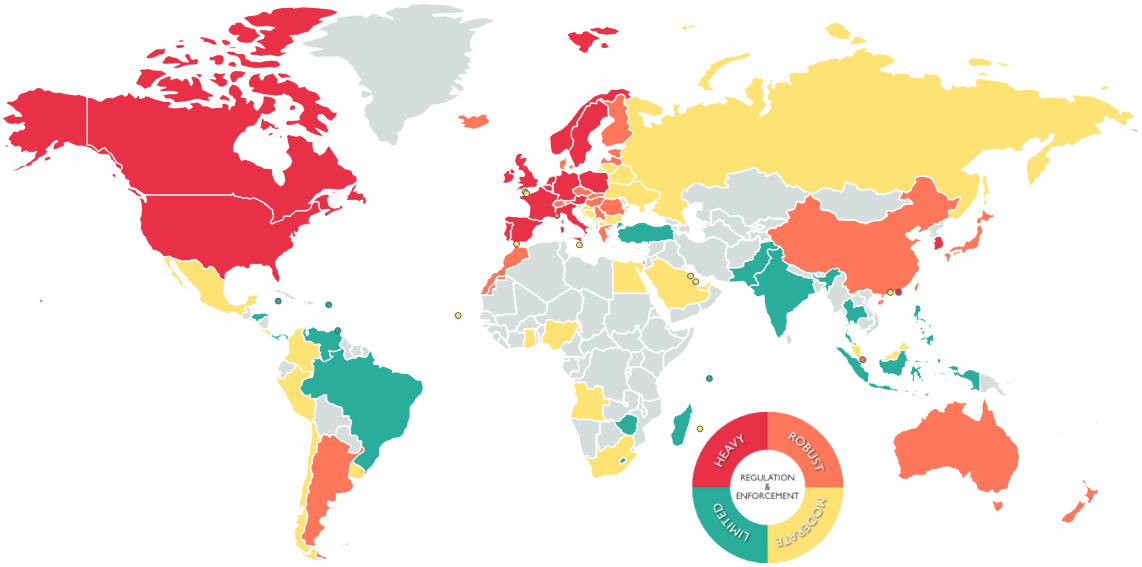

Figura 3.1. Nível de regulamentação das leis de proteção de dados ao redor do mundo (Fonte: [2]).

porânea, essa liberação pode colocar em risco a privacidade dos indivíduos. Se os dados coletados forem mantidos em seu formato original, ou se esses dados caírem em mãos erradas por meio de vazamento de informações, indivíduos podem ser facilmente identificados, e assim, ter sua privacidade violada. Quando dados são publicados, seja para qualquer tipo de análise, um usuário malicioso pode ser capaz de descobrir informações sensíveis sobre indivíduos por meio de seus semi-identificadores. Semi-identificadores são dados que podem ser combinados com informações externas e assim utilizados para reidentificar indivíduos. Logo, um adversário poderá descobrir que o registro no conjunto de dados publicado pertence a um indivíduo com uma probabilidade alta. Dessa forma, caso uma publicação aconteça de maneira ingênua, tal fato pode levar a sérios riscos de violação de privacidade, uma vez que esses dados fornecem informações sensíveis, tais como costumes sociais, doenças, preferências religiosas, sexuais, entre outras.

Para lidar com a privacidade de dados dos cidadãos, governos ao redor do mundo definem regulamentos obrigatórios os quais as organizações devem seguir, por exemplo, HIPAA (*Health Insurance Portability and Accountability Act*) [5] nos Estados Unidos, FIPPA (*Freedom of Information and Protection of Privacy Act*) [4] no Canadá, Diretivas de Proteção de Dados da União Europeia [3], entre outros. No Brasil, ainda não há uma consolidação de leis sobre privacidade nos moldes de países como Estados Unidos, Canadá, ou mesmo da União Europeia, todavia, o direito à privacidade no país é regido pela Constituição Federal de 1988, pelo Novo Código Civil (lei 10.406/02, em especial o artigo 21) e também pelo Marco Civil da Internet (lei número 12.965, sancionada em 23 de abril de 2014). A Figura 3.1 apresenta o nível de regulamentação das leis de proteção de dados ao redor do mundo. O objetivo dessas leis é controlar o acesso às informações que as organizações detêm o controle. Tais organizações também podem ter suas próprias políticas de privacidade. Mesmo assim, elas necessitam tomar medidas concretas para proteger os dados de seus usuários, investigando métodos e ferramentas para preservar dados confidenciais e não permitir que adversários violem sua privacidade.

Garantir a privacidade dos indivíduos tem impacto direto na qualidade dos dados publicados, visto que a privacidade e a utilidade dos dados são princípios inversamente proporcionais. Quanto maior a privacidade, menor será a utilidade dos dados para análise, e vice-versa. Manter a utilidade dos dados e simultaneamente garantir a privacidade dos indivíduos é um problema extremamente complexo. Por isso, vários modelos de privacidade de dados têm sido propostos por pesquisadores com o objetivo de resolver esta questão. Um modelo pode ser classificado em sintático ou diferencial. O primeiro refere-se a uma condição na qual os dados devem obedecer antes de serem disponibilizados. Ou seja, os dados só serão publicados se obedecerem a um determinado critério. Já o segundo propõe-se a disponibilizar resultados de consultas, tendo como base um modelo matemático onde são compartilhadas apenas informações estatísticas sobre o conjunto de dados original.

Este capítulo tem por objetivo aprofundar conhecimentos em torno do tema privacidade, particularmente sobre fundamentos e técnicas para assegurar que indivíduos não possam ser reidentificados, além de descrever aplicações da privacidade de dados no mundo real, que visam balancear utilidade e privacidade. A Seção 3.2 apresenta os princípios básicos sobre o tema, bem como a definição de privacidade e como ela pode ser garantida. A Seção 3.3 esclarece o problema da utilidade dos dados em oposição à garantia de privacidade. Já a Seção 3.4 apresenta uma visão geral das técnicas mais utilizadas para anonimizar dados, destacando a generalização, a supressão e a perturbação de dados. Métricas para determinar a qualidade dos dados anonimizados são apresentadas na Seção 3.5. Os modelos sintáticos de privacidade mais comuns são descritos na Seção 3.6. O modelo de privacidade diferencial, proposto nos últimos anos como novo paradigma de privacidade, é detalhado na Seção 3.7. A Seção 3.8 exemplifica a utilização das técnicas vistas neste capítulo por meio de aplicações reais. E por fim, a Seção 3.9 apresenta as considerações finais do capítulo.

3.2. Fundamentos da Privacidade de Dados

O estudo da privacidade abrange disciplinas desde a filosofia à ciência política, teoria política e legal, ciência da informação e, de forma crescente, engenharia e ciência da computação. Um consenso entre os pesquisadores é que privacidade é um assunto complexo, com muitas questões envolvidas. O conceito de privacidade está relacionado a pessoas, mais precisamente ao direito que as pessoas têm em manter um espaço pessoal, sem interferências de outras pessoas ou organizações. Dessa forma, compete a elas decidir manter suas informações sob seu exclusivo controle, ou informar, decidindo a quem, quando e onde suas informações estarão disponíveis. Contudo, a privacidade tem sido utilizada como moeda de troca em serviços "gratuitos", nos quais o usuário provê informações pessoais para fazer uso desses serviços. Nesse contexto, quais as informações que, de fato, não devem ser divulgadas? A que tipos de ataque minhas informações estão sujeitas? Quais as maneiras de proteger meus dados? Nesta seção iremos discutir essas e outras questões relativas aos fundamentos da privacidade de dados.

3.2.1. O que é Privacidade e por que ela é importante

Muitas definições de privacidade foram elaboradas e defendidas ao longo do tempo. O reconhecimento do direito do indivíduo à privacidade está enraizado na história. De acordo

com Laurant [29] a mais antiga referência à privacidade remonta ao Alcorão e às declarações de Maomé. Mais tarde, no século XIX, o conceito de privacidade foi estendido à aparência pessoal dos indivíduos, provérbios, atos, crenças, pensamentos, emoções, sensações, etc. Já na década de 1980, Gavison [23] define privacidade como uma condição medida em termos de grau de acesso que outros têm a você, com base na informação, atenção e proximidade. Dessa forma existem, em termos jurídicos, três princípios fundamentais e independentes que compõem a privacidade: o sigilo (ou segredo); o anonimato; o isolamento (ou solidão). Já na última década, Daniel Solove [40] propôs uma taxonomia de privacidade que consiste em quatro grupos de atividades: (i) coleta de informações; (ii) processamento de informações; (iii) divulgação de informações; (iv) invasão. Cada grupo contém uma variedade de atividades que podem criar problemas relacionados à privacidade.

Independentemente do tempo e do meio onde os indivíduos estão inseridos, existe a necessidade de privacidade, mesmo que alguém a desconheça ou não se importe com ela. A privacidade encontra uma barreira para a sua existência no mundo virtual, devido à facilidade da transmissão da informação. Quando conectado a uma rede, um dispositivo (computador, notebook, tablet, celular, etc.) pode estar vulnerável a todo tipo de ataque externo. As informações contidas nestes dispositivos podem ser acessadas e divulgadas para o resto da rede.

É bastante comum os usuários confundirem os conceitos de privacidade e de segurança. Apesar de privacidade e segurança serem temas relacionados, suas definições remetem a polos opostos, como duas pontas de uma gangorra. É inevitável que um concorra com o outro. Por exemplo, recentemente, mesmo diante do terrorismo que acomete os Estados Unidos, grande parte dos americanos afirmaram que não estão dispostos a compartilhar seus e-mails pessoais, mensagens de texto, telefonemas e registros de atividade na Web com investigadores anti-terrorismo [6]. Em outras palavras, eles não estão dispostos a abrir mão de sua privacidade em troca de segurança.

Quando se trata de dados, a segurança visa regular o acesso durante todo o ciclo de vida do dado, enquanto a privacidade define como será realizado esse acesso, na maioria das vezes com base em leis e políticas de privacidade. Neste ponto, também surge o conceito de controle de acesso como forma de fornecer segurança a um conjunto de dados. O controle de acesso se refere a regras específicas de quem está autorizado a acessar (ou não) determinados recursos, isto é, quando um conjunto de usuários está apto a acessar um conjunto de dados. A privacidade aqui está associada a regras de controle de acesso efetivas, que permitem a revelação da informação apenas por usuários autorizados. Contudo, a privacidade dos indivíduos não está garantida apenas com o controle de acesso eficiente, visto que os usuários com acesso àquelas informações podem ser maliciosos, e assim capazes de divulgar informações sensíveis acerca daqueles indivíduos.

Incidentes envolvendo violação de privacidade têm ocorrido em diversos lugares ao redor do planeta. Em 2014, milhares de funcionários do serviço de ambulância e requerentes de benefícios de programas de habitação do Reino Unido tiveram seus dados pessoais violados acidentalmente [7]. Informações como idade, gênero e religião de mais de 2.800 funcionários foram publicados na Web por engano. Outro fato foi divulgado pela *Community Health Systems* (CHS), nos Estados Unidos, a qual afirmou que cerca

de 4,5 milhões de dados de identificação de pacientes haviam sido roubados [1]. Neste caso, nenhum dado médico/clínico ou número de cartão de crédito foi violado, mas os dados incluíam informações que seriam úteis para descoberta de identidade, como nomes de pacientes, endereços, datas de nascimento, números de telefone e números de seguro social de milhões de indivíduos. Quando estes incidentes desfavoráveis acontecem, as organizações enfrentam processos legais, prejuízos financeiros, perda de imagem e, acima de tudo, a perda de seus clientes.

Cada vez mais organizações têm desembolsado milhões de dólares para proteger as informações de seus clientes. E por que gastar tanto recurso assim? É fato que algumas informações são tão importantes que não podem ser reveladas. Geralmente as informações que necessitam de maior privacidade e que não podem sofrer ataques são as informações que identificam unicamente indivíduos ou que são sensíveis a eles, como por exemplo condição financeira, doenças, etc. As características de cada uma dessas informações são detalhadas a seguir.

3.2.2. Microdados e Privacidade

Dados podem ser representados por uma tabela, onde cada coluna corresponde a um atributo e cada linha a um registro. Esses dados são denominados microdados (ou dados tabulados). Em geral, microdados têm um conjunto fixo de atributos, que são comuns em uma coleção de registros. Quatro tipos de atributos podem existir em conjuntos de dados desse tipo [20]:

- **Identificadores explícitos**: são atributos que identificam unicamente indivíduos, tais como "nome", "CPF", "e-mail", etc., e são sempre removidos antes de serem publicados;
- **Semi-identificadores**: são todos aqueles atributos que não são identificadores explícitos mas podem potencialmente identificar um indivíduo, especialmente quando agrupados. São exemplos de semi-identificadores em dados relacionais "data de nascimento" e "CEP";
- **Atributos sensíveis**: contém informações sensíveis sobre indivíduos, tais como "doença", "salário", etc.;
- **Atributos não sensíveis**: é qualquer tipo de atributo que não se enquadra em nenhuma das categorias anteriores.

Dados sensíveis armazenados em sistemas de banco de dados sofrem riscos de divulgação não autorizada. Por esse motivo, tais dados precisam ser protegidos. Para exemplificar os tipos de atributos, tendo em vista os riscos de divulgação não autorizada, considere o exemplo de microdados de duas tabelas, uma de cliente bancário e outra de conta. A Tabela 3.1 mostra dados de clientes contendo identificadores explícitos e semi-identificadores. A tabela considerada como tal, isto é, sem nenhuma modificação, não possui nada de tão confidencial, pois a maioria das informações nela contidas também estão disponíveis em bases de dados públicas ou em redes sociais, como por exemplo o Facebook.

Identificadores Explícitos		Semi-identificadores			
ID	**Nome**	**Idade**	**Gênero**	**Endereço**	**Telefone**
1	David	22	Masculino	Av. L	98533 1234
2	John	23	Masculino	Av. K	98772 2531
3	Helton	25	Masculino	Av. K	98156 0092
4	Maria	32	Feminino	Rua J	99913 9026

Tabela 3.1. Exemplos de identificadores explícitos e semi-identificadores em dados tabulados de clientes.

Diferente da Tabela 3.1, a Tabela 3.2 apresenta dados de contas bancárias de seus usuários contendo algumas informações consideradas sensíveis e que não devem ser divulgadas. Todavia, quando acessada de forma isolada, possui informações sobre seus clientes que são úteis para análises e descoberta de padrões.

Identificadores Explícitos	Atributos Sensíveis		
ID	**Conta**	**Tipo**	**Saldo (R$)**
1	2234-0	Corrente	1.033,25
2	7749-2	Corrente	814,92
3	8491-7	Corrente	515,09
4	5723-1	Poupança	2.194,79

Tabela 3.2. Exemplos de identificadores explícitos e atributos sensíveis em dados tabulados de conta bancária.

A partir da divulgação das Tabelas 3.1 e 3.2, adversários podem obter informações de uma vítima e serem capazes de explorar conhecimento sobre ela por meio de ligação de informações. Para isso, eles associam registros públicos a um indivíduo alvo, cujas informações estão contidas no conjunto de dados publicado, violando assim sua privacidade. Além disso, adversários também são capazes de inferir valores de atributos sensíveis a partir desse conhecimento.

3.2.3. Conhecimento Adversário e Ataques à Privacidade

Uma violação de privacidade ocorre por meio de um ataque, i.e., quando um adversário é capaz de associar o proprietário de um dado a um registro em um conjunto de dados, utilizando um conhecimento previamente adquirido de fontes externas. Por exemplo, o adversário pode saber que a vítima mora ao lado de sua residência, assim ele pode inferir informações como endereço, CEP, gênero da vítima, etc. O adversário pode também utilizar dados de serviços baseados em localização, como um *checkin* em uma rede social realizado por uma vítima em uma determinada localização. O adversário pode ainda ter acesso a dados abertos de uma vítima, caso ela seja funcionária de órgãos públicos, por exemplo. Dessa forma, o conhecimento adversário é tido muitas vezes como imprevisível e deve ser considerado em soluções de preservação de privacidade, mesmo diante da

incerteza de como ele foi obtido.

Um adversário é capaz de violar a privacidade de indivíduos através dos seguintes ataques:

- **Ataque de Ligação ao Registro**: o objetivo do adversário é reidentificar o registro de um indivíduo específico ou de um indivíduo qualquer, cujas informações aparecem no conjunto de dados publicado;
- **Ataque de Ligação ao Atributo**: nesse tipo de ataque o adversário pode ser capaz de inferir atributos sensíveis de uma vítima mesmo sem reidentificar seus registros, baseando-se no conjunto de valores sensíveis associados ao grupo no qual a vítima pertence;
- **Ataque de Ligação à Tabela**: tanto os ataques de ligação ao registro quanto de ligação ao atributo assumem que o atacante sabe que o registro da vítima foi publicado. Nesse tipo de ataque o adversário está interessado em inferir com convicção a presença ou a ausência da vítima nos dados publicados;
- **Ataque Probabilístico**: esse tipo de ataque não foca em quais registros, atributos ou tabelas o atacante pode associar informações sensíveis a indivíduos, mas sim destaca como o atacante mudaria seu pensamento probabilístico acerca de um indivíduo após ter acessado o conjunto de dados publicado.

3.2.4. Maneiras de Proteger Dados Sensíveis

Uma das tarefas mais árduas no tema de privacidade e segurança da informação é proteger dados sensíveis de possíveis ataques. Quando um conjunto de dados é disponibilizado para fins estatísticos, de pesquisa, ou de testes, técnicas de preservação de privacidade são necessárias para evitar a descoberta de informações sensíveis por usuários maliciosos. Por exemplo, um hospital disponibiliza publicamente dados sobre seus pacientes para auxiliar pesquisadores da área médica a descobrirem causas de doenças, ou para estatísticos afirmarem a frequência da ocorrência de um determinado vírus. Uma vez que esses dados contêm informações sensíveis sobre pacientes, tal hospital não deve liberá-los de uma maneira ingênua, devido ao alto risco de violação da privacidade. Como forma de proteger efetivamente a privacidade dos indivíduos, o detentor dos dados precisa garantir que eventuais descobertas de informações não ocorram no conjunto de dados disponibilizado.

Existem três grandes maneiras de contornar esse problema: criptografia; tokenização; anonimização. A criptografia é considerada uma das técnicas mais antigas para se proteger dados e, quando bem executada, se torna uma técnica bastante robusta no quesito privacidade. Essa técnica utiliza um algoritmo capaz de embaralhar matematicamente dados sensíveis, gerando substitutos ilegíveis. Esses substitutos podem ser transformados de volta, para seus valores originais, através da utilização de uma chave de acesso. Nesse caso, dados criptografados não são legíveis e consequentemente não são úteis para análises. Outro quesito acerca da técnica de criptografia é o gerenciamento da chave de acesso. Caso ela não seja bem controlada e caia em mãos erradas, há uma total perda de privacidade. A criptografia não é o foco deste capítulo, visto que não é amplamente utilizada no campo da privacidade.

Tokenização é uma técnica de proteção de dados utilizada principalmente quando empresas buscam proteger dados confidenciais já armazenados ou em movimentação para a nuvem. Essa técnica gera aleatoriamente um valor de *token* sem formatação específica a partir de um registro original e armazena o mapeamento desse *token* com seu respectivo valor original em uma base de dados. Dessa forma, *tokens* não podem ser revertidos aos seus valores originais sem o devido acesso à tabela de mapeamento. Por exemplo, o nome "Francisco" pode ser mapeado para o *token* "F+YCO" e o número de conta corrente "3256-6" pode se tornar "ES*X-2" após um processo de tokenização. A principal diferença entre essa técnica e a criptografia é que, na tokenização, os dados originais são completamente substituídos por caracteres que não tem nenhuma conexão com os dados originais. Outro fato relevante sobre essa técnica é que, embora o *token* seja utilizável dentro de seu ambiente de aplicação nativo, é completamente inútil em outro contexto. Dessa forma, a tokenização é ideal para proteger números de cartões de crédito, números de seguro social, além de outras informações que identificam indivíduos de forma única.

A anonimização é constituída por um conjunto de técnicas que modificam dados originais, de tal forma que os dados anonimizados não se assemelham aos dados originais, mas ambos possuem semântica e sintaxe bastante semelhantes. O termo anonimato representa o fato do sujeito não ser unicamente caracterizado dentro de um conjunto de sujeitos. O intuito da anonimização é poder compartilhar informações com outras entidades, as quais poderão utilizá-las para diversas finalidades, sem que haja violação de privacidade. Modificações implicam em perda de informação e, consequentemente, em diminuição da utilidade dos dados. Logo, o desafio da anonimização é transformar os dados de tal forma que a privacidade dos indivíduos é protegida, enquanto a utilidade dos dados é mantida. Esse desafio é discutido com mais detalhes a seguir.

3.3. O Desafio de Disponibilizar Dados

Não se deve apenas pensar na privacidade dos indivíduos e esquecer a utilidade dos mesmos. Dados disponibilizados precisam ser úteis para que outros tipos de usuários possam realizar atividades importantes. Isso torna o processo de disponibilização de dados uma tarefa nada trivial. Além disso, existem diversas partes envolvidas em um cenário de disponibilização de dados [44], conforme mostra a Figura 3.2. As características dessas partes são detalhadas a seguir:

- **Cliente / Fornecedor dos Dados**: é representado por um indivíduo ou por uma organização que compartilha seus próprios dados. Por exemplo, um indivíduo que fornece seus dados pessoais como nome, endereço, gênero, data de nascimento, telefone, e-mail para cadastro em uma agência bancária;

- **Organização**: é evidenciada por qualquer tipo de entidade que serve à realização de ações de interesse social, político, etc. Por exemplo organizações bancárias, de saúde, de comércio eletrônico, ou mesmo de redes sociais, que detém uma certa quantidade de informações sobre seus clientes. Elas são responsáveis por proteger os dados de seus usuários a qualquer custo. Em caso de vazamento de informações, as organizações enfrentam prejuízos financeiros, processos legais, perda de reputação e perda de seus clientes;

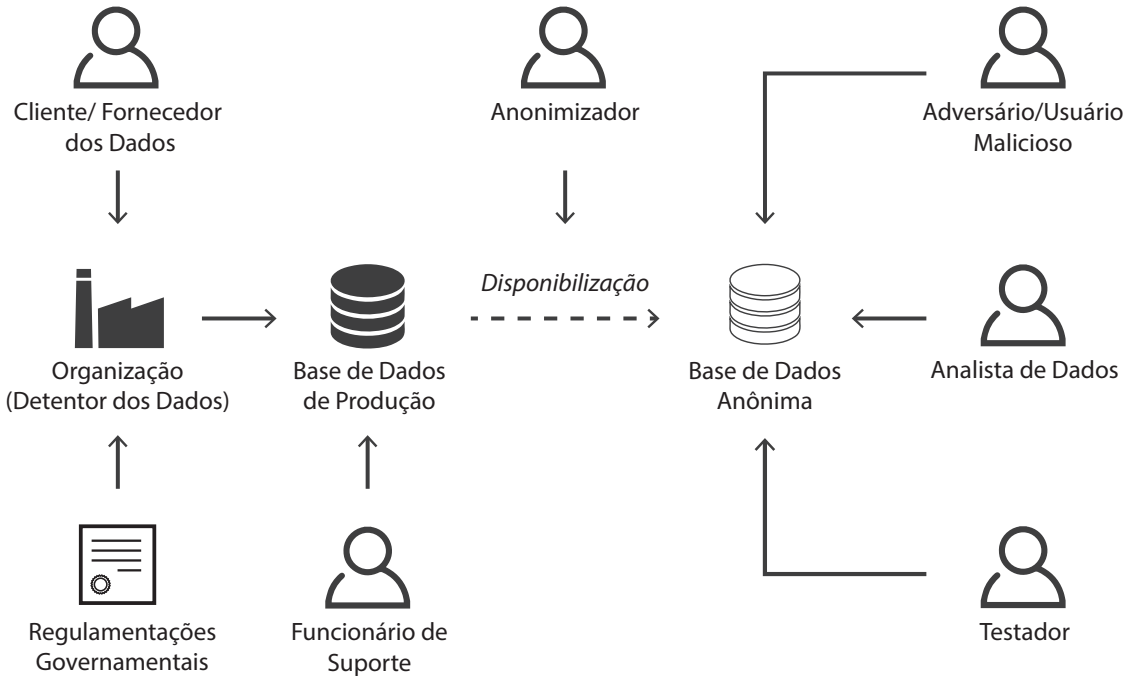

Figura 3.2. Partes envolvidas em um cenário de disponibilização de dados.

- **Regulamentações Governamentais**: define quais regras de proteção de dados as organizações devem seguir. HIPAA nos Estados Unidos e FIPPA no Canadá são exemplos de regulamentações. Vale ressaltar que as próprias organizações podem definir regulamentações internas;

- **Anonimizador**: é representado por um indivíduo ou organização que aplica técnicas de anonimização sobre os dados, para que analistas e testadores possam usufruir dos dados sem comprometer a privacidade dos mesmos. O papel do anonimizador também pode ser representado por um software que realiza a transformação dos dados de maneira automática.

- **Analista de dados**: utiliza os dados anonimizados para realizar minerações e descobrir padrões. Algumas regulamentações determinam que análises só podem ser realizadas sobre dados anonimizados. Por esse motivo, é importante que os dados disponibilizados suportem as funcionalidades de mineração de dados;

- **Testador**: a terceirização de testes de software é comum entre muitas empresas. Testes de alta qualidade exigem dados de teste também com alta qualidade, que estão presentes nos sistemas de produção e contém informações sensíveis de clientes. Para que testes possam ser realizados com eficiência, o testador precisa de dados extraídos dos sistemas de produção, porém anonimizados e provisionados para testes;

- **Funcionário de Suporte**: tem como objetivo auxiliar no funcionamento dos requisitos de negócio dos clientes. Por esse motivo, esse colaborador possui acesso ao conjunto de dados original e a todos os dados sensíveis de clientes;

- **Adversário**: também conhecido como atacante ou usuário malicioso. Ele visa obter dados acerca de um indivíduo específico e assim descobrir informações sensíveis que violem sua privacidade.

Diante de todas essas partes envolvidas no processo de disponibilização de dados, um questionamento pode ser levantado: como proteger a privacidade dos clientes/fornecedores de dados e ao mesmo tempo garantir que os dados sejam úteis para testadores ou analistas de dados?

Pesquisas mostraram que a abordagem mais promissora para proteger a privacidade dos indivíduos é anonimizar os dados antes de disponibilizá-los publicamente ou para terceiros [21, 47]. Para isso, o detentor dos dados deve modificá-los de tal forma que nenhuma informação sensível sobre indivíduos possa ser descoberta a partir de uma publicação. Além disso, ele deve garantir que os dados sejam úteis para que eventuais análises possam ser efetuadas com qualidade. Por esse motivo, o detentor dos dados também deve buscar uma solução que preserve ao máximo a utilidade das informações as quais ele deseja disponibilizar. Em contrapartida, a não disponibilização de dados impede que governos, organizações e instituições possam tirar proveito de análises importantes, padrões e tendências para a sociedade, pesquisas científicas, entre outras atividades, dificultando assim o crescimento de tais entidades. Portanto, quanto maior o grau de privacidade associado aos dados, menos úteis aqueles dados serão para eventuais análises.

É fato que este é um problema desafiador, uma vez que qualquer alteração sobre os dados distorce sua utilidade. Além disso, nem sempre regulamentações são favoráveis no balanceamento entre privacidade e utilidade. Para demonstrar essa questão, considere o seguinte caso no domínio de saúde: a regulamentação governamental HIPAA, nos Estados Unidos, afirma que, qualquer atributo pessoalmente identificável (por exemplo nome, telefone, data de internação, etc.), deverá ser completamente anonimizado no conjunto de dados publicado. Ao mesmo tempo, profissionais de saúde podem compartilhar dados de paciente com parceiros externos para que sejam realizadas análises da eficácia de um determinado tratamento, por exemplo. Contudo, será impossível analisar este acontecimento, visto que a data de registro do paciente é anonimizada de acordo com as leis de privacidade da HIPAA, ou seja, não é possível analisar a eficácia do tratamento sem ter conhecimento sobre o tempo em que o paciente está sendo acompanhado. Dessa forma, há uma enorme demanda entre pesquisadores na área de privacidade em propor modelos, técnicas e algoritmos que atendam aos mais diversos tipos de balanceamento entre privacidade e utilidade em inúmeros contextos. A Figura 3.3 mostra o caso geral de balanceamento entre utilidade e privacidade [44].

A estratégia de criptografia citada na seção anterior, não fornece utilidade (valor 0), porém dispõe de alta privacidade (valor 1) quando os dados estão criptografados. De maneira oposta, ela fornece alta utilidade (valor 1) mas nenhuma privacidade (valor 0) quando os dados estão descriptografados. Ou seja, é uma estratégia praticamente 0 ou 1 em termos de privacidade e utilidade. Já na anonimização de dados, o nível de privacidade e utilidade pode flutuar entre o intervalo $[0, 1]$, idealmente na zona cinza da Figura 3.3. Dessa maneira, pode-se controlar o nível de ambos os indicadores. Assim, a melhor solução de balanceamento entre privacidade e utilidade dependerá da métrica de utilidade e do modelo de privacidade adotados para a disponibilização de dados.

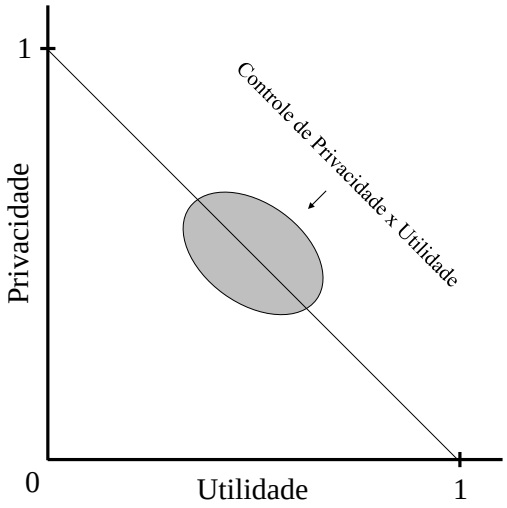

Figura 3.3. Balanceamento entre utilidade e privacidade.

3.4. Técnicas de Anonimização de Dados

A publicação de dados pode levar a sérios riscos de violação de privacidade devido à existência dos semi-identificadores. Isso pode acarretar em consequências graves por causa do uso não autorizado de informações sensíveis pertencentes aos indivíduos. Como forma de solucionar esse problema, uma estratégia ingênua seria a não publicação dos dados, para qualquer finalidade [48]. Contudo, isso evitaria que governos, organizações, etc. pudessem tirar proveito de análises importantes de padrões e tendências para a sociedade, dificultando o possível crescimento dessas entidades. Outra maneira de evitar o problema da publicação de dados seria disponibilizar apenas dados estatísticos para análise, porém essa estratégia é limitada ao conhecimento estatístico que o detentor dos dados possui, uma vez que ele deseja apenas publicar o dado, e não o analisar previamente antes de liberá-lo. A abordagem mais promissora para solucionar o problema da preservação de privacidade em uma publicação é anonimizar os dados antes de qualquer liberação [21]. Uma abordagem convencional para anonimizar dados tem sido praticada com a remoção dos identificadores explícitos de indivíduos, como nome, CPF, e-mail, etc. do conjunto de dados antes de uma publicação. Contudo, o trabalho em [42] demonstra que, simplesmente remover esses identificadores, não é suficiente para proteger a privacidade dos indivíduos, devido à existência dos semi-identificadores.

Esta seção apresenta uma visão geral das técnicas mais utilizadas para anonimização e publicação de dados. Em um processo de anonimização, um conjunto de dados original D é transformado em um novo conjunto D', por meio de modificações. O objetivo é evitar a descoberta de informações sensíveis por usuários maliciosos. As seguintes técnicas de anonimização são detalhadas a seguir: *generalização*; *supressão*; *perturbação*. Todas elas produzem um conjunto de dados D' menos preciso que o conjunto original D, no entanto ambos os conjuntos diferem no quesito perda de informação e também na proteção da privacidade ocasionada por cada técnica.

3.4.1. Generalização

O objetivo da generalização é aumentar a incerteza de um adversário ao tentar associar um indivíduo a seu registro, ou a informações sensíveis, no conjunto de dados publicado. Nesta técnica, os valores dos atributos que são considerados semi-identificadores são substituídos por valores semanticamente semelhantes, porém menos específicos. Dessa forma, a generalização preserva a veracidade dos dados quando aplicada sobre registros. A técnica de generalização pode ser aplicada em atributos tanto categóricos quanto numéricos. Assim, para cada categoria de atributos, pode existir uma hierarquia de generalização que representa a semântica desses atributos. Ao utilizar essa hierarquia, os valores de registros de uma determinada categoria são substituídos por valores menos específicos, abrangendo assim um maior domínio de valores para aquele atributo. Por exemplo, o valor tabulado "25" (domínio numérico) pode ser substituído pelo intervalo "[20, 29]" em um conjunto de dados publicado. A operação contrária à generalização é denominada especialização.

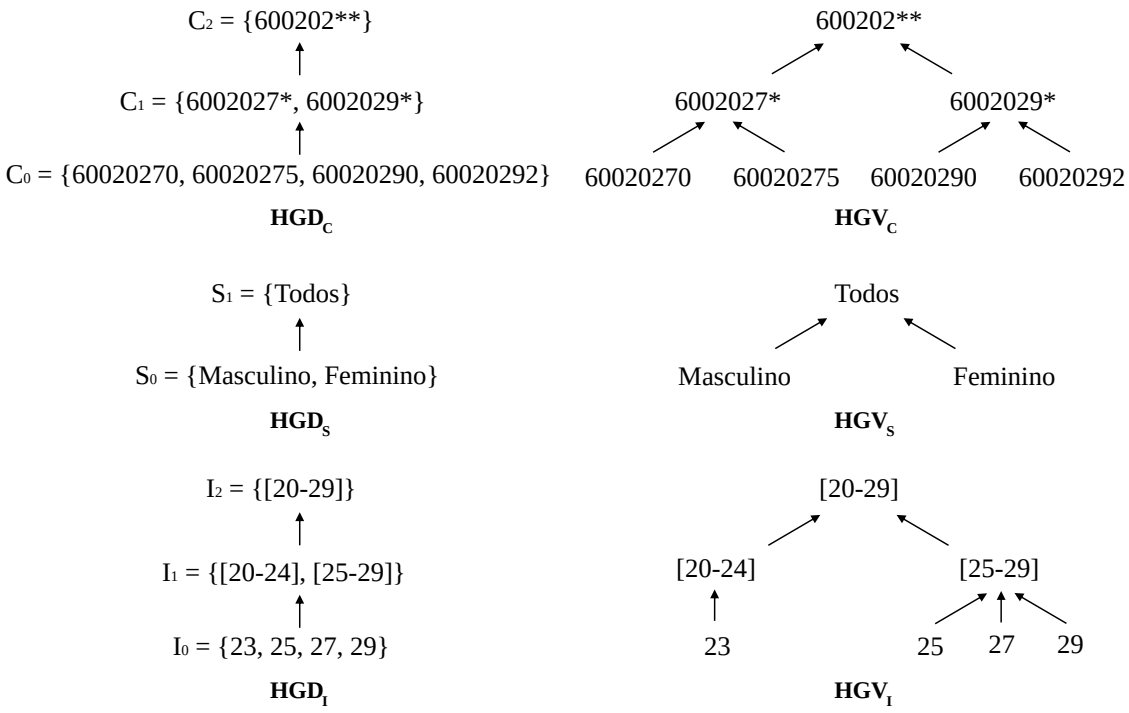

Figura 3.4. Exemplos de HGD e HGV para os atributos CEP (C), gênero (S) e idade (I).

Uma hierarquia de generalização pode ser predefinida por um domínio, sendo assim chamada de *hierarquia de generalização de domínio* (HGD). Uma HGD consiste em um conjunto de domínios totalmente ordenados pelo domínio da generalização de um atributo. Intuitivamente, considere dois domínios Dom_i e Dom_j. A comparação $Dom_i <_D Dom_j$ indica que os valores em Dom_j são generalizações dos valores em Dom_i. Dessa maneira, dado um domínio *Dom* e uma HGD, pode-se definir uma *hierarquia de generalização de valor* (HGV) como sendo o conjunto de valores em um determinado domínio, mas que são parcialmente ordenados pelo domínio da generalização de um atributo. Dessa forma, sendo dois valores v_i e v_j, a comparação $v_i \leq_V v_j$ indica que v_j é

uma generalização de v_i. A Figura 3.4 apresenta exemplos de hierarquias de generalização de domínio e de valor para os atributos CEP, gênero e idade, respectivamente. Neste exemplo, considerando o atributo CEP, a seguinte comparação pode ser considerada: '60020270' $<_V$ '6002027*' $<_V$ '600202**'. Portanto, o valor '600202**' é uma generalização de '6002027*', que por sua vez também é uma generalização do valor '60020270'.

Operações de generalização aplicadas de maneira ingênua sobre atributos semi-identificadores podem não gerar dados úteis para eventuais análises. Por esse motivo, é necessário encontrar uma generalização mínima, isto é, o conjunto mínimo necessário de alterações que devem ser aplicadas a um conjunto de dados, com o objetivo de manter sua utilidade e ao mesmo tempo atender aos requisitos de privacidade estabelecidos. O conjunto de todas as generalizações possíveis para todos os atributos semi-identificadores formam uma estrutura de reticulado, onde cada nó dessa estrutura corresponde a uma possível estratégia de generalização. A solução ótima para o problema da generalização mínima pode ser dada pelo nó do reticulado que satisfaz os requisitos de privacidade e resulta na menor perda de informação. Uma abordagem para encontrar a solução ótima para esse problema seria enumerar todos os nós da estrutura de reticulado e retornar aquele que produz a menor quantidade de distorção nos dados. Contudo, alguns autores já provaram que o problema de encontrar a generalização ótima é NP-difícil [9, 49]. Logo, heurísticas devem ser utilizadas para reduzir o espaço de busca e encontrar uma solução aproximada da solução ótima. A Figura 3.5 mostra o conjunto de todas as generalizações possíveis para os atributos semi-identificadores CEP, gênero e idade, formando o retículo desses atributos. Cada elemento C_i, S_j e I_k representam o grau de generalização baseado em suas respectivas hierarquias de generalização.

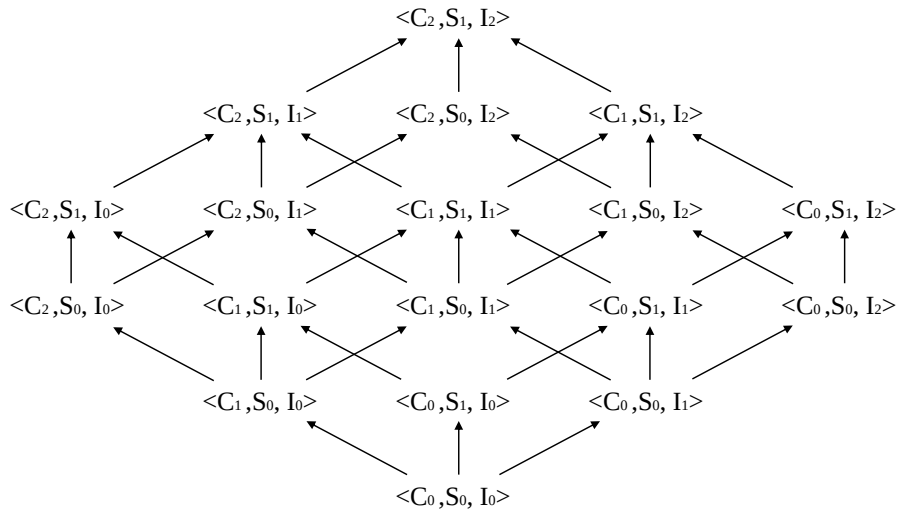

Figura 3.5. Reticulado do conjunto de todas as possíveis generalizações para os atributos CEP (C), gênero (S) e idade (I) baseado em suas respectivas hierarquias.

A operação de generalização pode ser aplicada tanto para todos os registros de um atributo semi-identificador quanto para apenas alguns. Como abordagem mais habitual, todos os registros de um atributo são mapeados para um mesmo valor generalizado (me-

nos específico), obedecendo sua hierarquia de generalização. Este processo é denominado *generalização global* [26, 31]. Por outro lado, diferentes registros do mesmo atributo podem ser generalizados com diferentes valores em uma hierarquia de generalização. Este processo é chamado de *generalização local* [24, 32]. Em resumo, a generalização global anonimiza todos os registros de um atributo da mesma maneira, utilizando sempre os mesmos valores da hierarquia de generalização, enquanto a generalização local anonimiza diferentes registros de um atributo com diferentes valores, obedecendo também a sua hierarquia de generalização.

Em se tratando de generalização a nível de atributo, pode-se aplicar essa técnica de quatro maneiras distintas: (i) generalização de domínio completo; (ii) generalização de sub-árvore; (iii) generalização entre irmãos; (iv) generalização de células. Os exemplos a seguir utilizarão a hierarquia de generalização de valor para o atributo *profissão*, conforme a Figura 3.6.

- **Generalização de domínio completo**: nesta abordagem, todos os valores de um atributo semi-identificador são generalizados para o mesmo nível, obedecendo sua hierarquia de generalização [32]. Esta técnica possui o menor espaço de busca quando comparada às outras técnicas, entretanto é a que produz mais distorção nos dados devido a exigência de todos os valores estarem anonimizados no mesmo nível na hierarquia de generalização. Por exemplo, se as profissões "Dentista" e "Médico" forem generalizadas para profissionais da "Saúde", então "Analista" e "Gerente" também devem ser generalizados para profissionais da "TI".

- **Generalização de sub-árvore**: neste esquema, a exigência de que todos os valores de um atributo estejam no mesmo nível da hierarquia de generalização aplica-se apenas aos valores das sub-árvore [25]. Em outras palavras, quando um valor de um atributo é substituído por um valor generalizado, obedecendo sua hierarquia de generalização, todos os outros valores nas folhas da sub-árvore também serão igualmente generalizados. Por exemplo, se a profissão "Médico" for generalizada para profissional da "Saúde", o valor "Dentista" também deverá ser generalizado, contudo "Analista" e "Gerente" não sofrem distorções, pois pertencem a outra sub-árvore na hierarquia de generalização.

- **Generalização de irmãos**: conhecido na literatura como *sibling generalization*, esta abordagem é semelhante à generalização de sub-árvore mas não exige que todos os valores de irmãos nas folhas de uma sub-árvore sejam generalizados [31]. Este esquema produz menos distorção do que o esquema de generalização de sub-árvore, pois atua apenas sobre os valores de atributos que necessitam ser distorcidos. No exemplo da Figura 3.6, caso a profissão "Analista" seja generalizada para um profissional da "TI", os valores do conjunto de dados que possuem "Gerente" como profissão não necessitam de generalização.

- **Generalização de células**: nas abordagens anteriores, caso o valor de um atributo seja generalizado, então todos os demais registros com o mesmo valor também devem ser generalizados. Dessa forma, tais abordagens são consideradas generalizações globais, anonimizando todos os registros de um atributo da mesma maneira. Já

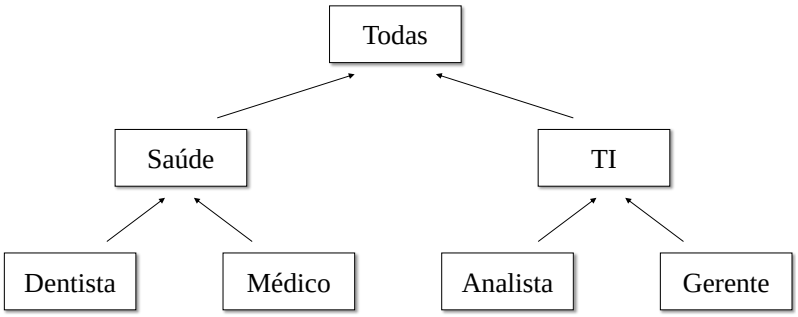

Figura 3.6. Hierarquia de generalização para o atributo semi-identificador Profissão.

no esquema de generalização de células, apenas algumas instâncias de um atributo semi-identificador são generalizadas, permitindo que outras instâncias se mantenham inalteradas [45].

3.4.2. Supressão

A supressão de dados é outra estratégia utilizada para compartilhamento e publicação de dados que ao mesmo tempo preserva a veracidade dos mesmos. Essa é uma técnica na qual um ou mais valores em um conjunto de dados são removidos ou substituídos por algum valor especial, possibilitando a não descoberta de semi-identificadores por adversários. A supressão de dados também pode ser vista como um tipo específico de generalização, no qual os registros são generalizados para o valor menos específico (nó raiz) na hierarquia de generalização de valor, que engloba todos os valores de um determinado atributo [48].

De maneira semelhante à técnica de generalização, a supressão de dados pode ser aplicada de maneira global ou local. A *supressão global* refere-se à remoção de todas as instâncias de um valor de atributo, garantindo que aqueles valores não serão descobertos em um conjunto de dados publicado, uma vez que todos foram removidos. Já a *supressão local* é caracterizada pela remoção de apenas algumas instâncias de um valor de atributo, contudo deve-se garantir que os valores restantes não possam ser descobertos. Os principais tipos de supressão de dados são:

- **Supressão de registro**: nessa abordagem, um registro é removido inteiramente do conjunto de dados. Consequentemente, nenhum valor de atributo é disponibilizado para os usuários [26, 31].
- **Supressão de valor**: refere-se a remoção ou a substituição de todas as instâncias de um valor de um atributo por um valor especial (como "*" ou "todos"). Por exemplo, os valores de atributo salário abaixo de R$ 30.000,00, em uma tabela de empregados, podem ser removidos ou substituídos por "*", enquanto os demais valores não sofrem distorções [46, 48].
- **Supressão de células**: nessa técnica, apenas algumas instâncias de valores de um atributo são removidas ou substituídas por um valor especial, caracterizando uma *supressão local* [36]. Por exemplo, pode-se remover apenas metade dos valores de

atributo salário abaixo de R$ 30.000,00, em uma tabela de empregados. Assim, instâncias de salário podem conter valores abaixo ou acima de R$ 30.000,00, além de valores suprimidos. Essa estratégia pode levar a inconsistências em eventuais análises de dados.

3.4.3. Perturbação

Essa abordagem tem sido comumente utilizada em controle de descoberta estatística [25], devido à sua simplicidade, eficiência e capacidade de preservar informações estatísticas. A ideia geral dessa técnica é substituir os valores dos atributos semi-identificadores originais por valores fictícios, de modo que informações estatísticas calculadas a partir dos dados originais não se diferenciem significativamente de informações estatísticas calculadas anteriormente sobre os dados perturbados. Ao contrário das técnicas de generalização e supressão, que preservam a veracidade dos dados, a perturbação resulta em um conjunto de dados com valores sintéticos. Muitas vezes isso acarreta em informações sem sentido para aqueles que irão utilizá-las. As técnicas mais comuns para a perturbação de dados são:

- **Adição de ruído**: essa técnica é geralmente aplicada sobre atributos numéricos. A ideia geral é substituir um valor original de atributo "v" por "$v + r$", onde "r" é um valor, denominado ruído, escolhido aleatoriamente a partir de uma distribuição. Um valor de atributo "v" também pode ser substituído pelo produto "$v \times r$" Em outras palavras, os valores de atributos são perturbados com um determinado nível de ruído, que pode ser adicionado ou multiplicado pelo valor original de cada atributo [43]. Consequentemente, mesmo que um atacante consiga identificar um valor individual de um atributo confidencial, o valor original não será revelado. A vantagem dessa técnica é que ela preserva algumas propriedades estatísticas, como média e correlação, mas ao mesmo tempo ela pode gerar alguns valores sem significado ou sem expressividade;

- **Permutação de Dados**: nessa abordagem, dois valores do mesmo atributo (de dois registros diferentes) são permutados. Isso mantém algumas características estatísticas dos dados, como contagem e frequência dos atributos [16]. Essa técnica não altera o domínio dos atributos, todavia as possíveis permutações de valores para diferentes atributos podem gerar registros sem sentido, e consequentemente, informações equivocadas. Por exemplo, considere os atributos "gênero" e "profissão". Caso haja uma permutação entre valores que gere um registro contendo "masculino" e "professora", esse registro não faria sentido, visto que o valor "professora" é do gênero feminino;

- **Geração de Dados Sintéticos**: nessa técnica, primeiramente um modelo estatístico é gerado a partir do conjunto de dados e, em seguida, são gerados dados sintéticos que seguem tal modelo [8]. São esses dados sintéticos que devem ser disponibilizados para o usuário final. A vantagem dessa técnica é que todas as propriedades estatísticas dos dados são preservadas. Contudo, pode-se gerar também alguns valores sem sentido e que não existem no mundo real.

O mascaramento de dados também é um método de perturbação utilizado para publicar conjuntos de dados com informações que pareçam reais, mas que não revelam informações sobre nenhum indivíduo. O objetivo principal do mascaramento é disponibilizar bases de dados para testes ou treinamento de usuários. Isso protege a privacidade dos dados pessoais presentes no banco de dados, bem como outras informações sensíveis que não possam ser colocadas à disposição para um time de testes ou para usuários em treinamento. Algumas técnicas de mascaramento de dados são descritas a seguir:

- **Substituição**: corresponde à substituição aleatória de dados por informações similares, mas sem nenhuma correlação com o dado real. Por exemplo, a substituição de sobrenome de família por algum outro proveniente de uma lista aleatória de sobrenomes;
- **Embaralhamento (*Shuffling*)**: é uma estratégia semelhante a substituição mas com a diferença de que o dado é derivado da própria coluna da tabela. Assim, o valor do atributo A em uma determinada tupla c_i é substituído pelo valor do atributo A em uma outra tupla c_k, selecionada randomicamente, onde $i \neq k$;
- ***Blurring***: essa técnica é aplicada a dados numéricos e datas. A técnica altera o valor do dado por alguma percentagem aleatória do seu valor real. Logo, pode-se alterar uma determinada data somando ou diminuindo um determinado número de dias, de forma aleatória;
- **Anulação/Truncagem**: essa técnica substitui os dados sensíveis por valores nulos (NULL) no conjunto de dados publicado. Geralmente esta técnica é utilizada quando os dados existentes na tabela não são necessários à realização de testes ou treinamentos.

3.5. Perda de Informação e Medidas de Utilidade dos Dados

A anonimização de dados causa perda de informação e muitas vezes compromete a utilidade dos dados, ou seja, quanto mais anonimizados forem os dados, menos úteis eles serão para o usuário final. Como forma de preservar a utilidade dos dados publicados, deve-se assegurar que o mínimo de distorção deva ser gerado na anonimização. Essa distorção causada por um processo de anonimização é denominada perda de informação. Há algumas métricas para se medir a perda de informação. Tais métricas podem ser utilizadas tanto para medir a utilidade do conjunto de dados publicado em relação aos dados originais, ou serem utilizadas como métricas de busca, com o objetivo de guiar os passos em busca da melhor solução de anonimização no espaço de todas as possibilidades [20].

A seleção de uma métrica de utilidade apropriada depende principalmente do algoritmo de anonimização adotado e do objetivo da disponibilização de dados. Por exemplo, se um algoritmo de anonimização utiliza uma hierarquia de generalização, então as métricas de perda de informação, que levam em conta o custo das operações de generalização, devem ser consideradas. Existem métricas que são utilizadas para cenários de publicação e disponibilização de dados em geral. Essas métricas, denominadas *métricas de uso geral*, são utilizadas no cenário em que o publicador dos dados não tem conhecimento prévio

sobre a área de aplicação dos dados a serem liberados ou quando não há objetivos específicos pré-definidos para o uso desses dados [20]. Por esse motivo, essas métricas devem levar em consideração muitos fatores e reter a maior quantidade de informação sempre que possível. Nesse caso, os dados publicados se tornam disponíveis para todos, como por exemplo na Internet, de modo que usuários com interesses distintos possam realizar análises de acordo com sua necessidade.

É fato que toda operação de generalização ou supressão causa alguma distorção sobre os dados. Em função disso, uma das métricas de uso geral utilizada para avaliar a utilidade dos dados é a *Precisão* [41]. Essa métrica obtém a distorção de dados por meio da atribuição de uma penalidade a cada instância de um valor de atributo que é generalizado ou suprimido. Se o valor de um atributo em um registro não é generalizado ou suprimido então não há distorção. Uma distorção é calculada para cada célula e é definida como a altura do nó correspondente ao valor generalizado na hierarquia de generalização dividida pela altura máxima da hierarquia. A distorção total do conjunto de dados é calculada pela soma de todas as distorções das células. Ao se obter a soma de todas as distorções das células e normalizar o resultado pelo número total de células, obtêm-se a métrica de precisão. Formalmente, a precisão de um conjunto de dados anonimizado $D(A_1,...,A_{N_a})$ é definida como:

$$Prec(D) = 1 - \frac{\sum_{i=1}^{N_a}\sum_{j=1}^{|D|}\frac{h}{|HGV_{A_i}|}}{|D| * |N_a|}$$

A precisão $Prec(D)$ assume valores entre 0 e 1. $|N_a|$ representa o número de atributos pertencentes ao conjunto de semi-identificadores, $|D|$ equivale ao número de registros da tabela, enquanto que h representa a altura da hierarquia de generalização de valor do atributo A_i após a generalização e por fim $|HGV_{A_i}|$ é a altura máxima da hierarquia. Quanto maior a precisão, maior a utilidade dos dados e consequentemente, os dados anonimizados são mais semelhantes ao conjunto de dados original.

A perda de informação causada por generalizações também pode ser medida utilizando a métrica *ILoss* [50]. Essa métrica captura a fração de nós folha, baseados em uma hierarquia de generalização, que são generalizados. Para um determinado registro, a métrica *ILoss* é calculada encontrando a soma dos valores de *ILoss* para todos os atributos daquele registro. É importante salientar que diferentes pesos podem ser aplicados a diferentes atributos. O *ILoss* global de um conjunto de dados pode ser obtido pela soma de todos os valores *ILoss* obtidos para os registros. Seja V_g um nó na hierarquia de generalização H de um atributo $A \in SI$, onde SI é conjunto de semi-identificadores. $|V_g|$ é definido como o número de folhas na sub-árvore de V_g. Seja $|D_A|$ o número de valores no domínio do atributo A, isto é, o número total de folhas de H. A métrica *ILoss* para um valor específico é calculada como:

$$ILoss(V_g) = \frac{|V_g| - 1}{|D_A|}$$

Considere agora $|W_i|$ como uma constante positiva que especifica a penalidade do atributo A_i sobre V_g. Essa constante pode ser definida pelo usuário a fim de determinar a

importância de cada atributo. A métrica $ILoss(r)$ em termos de registro é dada por:

$$ILoss(r) = \sum_{V_g \in r} (W_i * ILoss(V_g))$$

Finalmente, a perda total de informação no conjunto de dados anonimizado D é dada pela soma de todos os *ILoss* de registros:

$$ILoss(D) = \frac{\sum_{r \in D} ILoss(r)}{|D|}$$

Algumas métricas utilizam o conceito de classe de equivalência para mensurar a qualidade dos dados anonimizados. Formalmente, uma classe de equivalência é definida da seguinte forma:

Definição 1 *Considere uma série de atributos $A = \{A_1, ..., A_n\}$ em um conjunto de dados D. Uma classe de equivalência E é um conjunto de todos os registros em D que contém valores idênticos para os atributos em A.*

Por exemplo, a Tabela 3.3 abrange duas classes de equivalência para o conjunto de atributos $A = \{$Idade, Gênero, CEP$\}$. São elas: $E_1 = \{[20,30]$, Masculino, 60800***$\}$ e $E_2 = \{> 40$, *, 60790***$\}$.

Idade	Gênero	CEP
[20,30]	Masculino	60800***
[20-30]	Masculino	60800***
>40	*	60790***
>40	*	60790***
>40	*	60790***

Tabela 3.3. Conjunto de dados contendo duas classes de equivalência.

Uma das métricas que utiliza o conceito de classe de equivalência é a métrica de *discernibilidade*. Ela foi introduzida por Bayardo [26] e aborda a noção de perda de informação através de uma penalidade para cada registro por ser indistinguível de outros registros no conjunto de semi-identificadores. Nessa métrica, a qualidade é baseada justamente no tamanho da classe de equivalência E e no conjunto de dados D.

A métrica de discernibilidade C_{DM} atribui para cada registro r no conjunto de dados D uma penalidade determinada pelo tamanho da classe de equivalência contendo r. Se um registro pertence a uma classe equivalente de tamanho s, a penalidade para o registro é s. Se uma tupla é suprimida, então é atribuída uma penalidade de valor $|D|$. Essa penalidade refere-se ao fato de que uma tupla suprimida não pode ser distinguida de qualquer outra tupla no conjunto de dados. Formalmente, a discernibilidade é medida como sendo:

$$C_{DM} = \sum_{classesEq} |E|^2$$

No entanto, uma vez que essa métrica é baseada na dimensão das classes de equivalência, a mesma perda de informação é dada para todos os registos nas classes de equivalência de mesma dimensão. Porém, esses registros podem ser generalizados de maneiras diferentes e, portanto, possuem diferentes níveis de distorção que não são levados em conta pelo C_{DM}.

Outra métrica utilizada para mensurar a utilidade dos dados e que também lida com classes de equivalência é denominada *tamanho médio das classe de equivalência* (C_{AVG}) [32]. Esta métrica mede o quão bem uma partição, isto é, uma classe de equivalência, se aproxima do melhor caso. Seu objetivo é reduzir a média normalizada do tamanho das partições. Dado que cada registro é generalizado em classes de equivalência de pelo menos k registros indistinguíveis, essa métrica é definida como:

$$C_{AVG} = (\frac{totalRegistros}{totalClassesEq})/(k)$$

Utilizar a mesma métrica para disponibilizar dados a usuários distintos nem sempre é uma boa ideia, visto que essas métricas podem não se adequar as diferentes demandas dos diferentes usuários. Para contornar essa questão, são propostas as *métricas de finalidade específica*, que tem por objetivo atender as demandas de usuários quando um conjunto de dados é publicado para finalidades exclusivas. Desse modo, tais métricas são utilizadas caso a finalidade dos dados seja conhecida no momento da publicação ou caso os dados estejam sendo publicados para fins de mineração específicos. Nestes casos, as métricas de perda de informação que melhores se adequam aos objetivos específicos são justamente as métricas que devem ser adotadas antes da disponibilização. Por exemplo, se os dados forem liberados com o propósito de construir um classificador para um certo atributo, os valores que são importantes para a classificação não devem ser generalizados ou suprimidos. Ou seja, não se deve anonimizar os valores cujas distinções são essenciais para diferenciar os rótulos das classes do atributo alvo. Para atingir esse objetivo, o erro de classificação nas instâncias futuras deve ser considerado no cálculo da perda de informação.

A métrica mais comum utilizada nesta categoria é a *medida de classificação (CM)* [25]. A ideia dessa métrica é penalizar cada registro r que é suprimido ou generalizado para uma classe de equivalência, em que a classe de r não é a classe majoritária. A classe majoritária é aquela que contém o maior número de registros. Dessa forma, CM pode ser calculado pela soma de todas as penalidades de cada registro, normalizado pelo número total de registros.

$$CM = \frac{\sum_{r \in D} Penalidade(r)}{N}$$

D é o conjunto de dados, N é o número de registros em D e $Penalidade(r)$ é definido como:

$$Penalidade(r) = \begin{cases} 1 & \text{se } r \text{ é suprimido} \\ 1 & \text{se } classe(r) \neq Maj(classeEq(r)) \\ 0 & \text{caso contrário} \end{cases}$$

Também existem algumas métricas que, além de informações, levam em consideração os requisitos de privacidade para se medir a perda de informação proveniente do processo de anonimização. O objetivo dessas métricas é buscar a anonimização que minimiza a perda de informação enquanto maximiza o ganho de privacidade. Tais métricas são denominadas *métricas de trade-off*. Dessa forma, essas métricas calculam tanto o ganho de informações quanto a perda de privacidade a cada iteração de anonimização, de modo que o *trade-off* ideal possa ser encontrado para ambos os requisitos necessários.

Suponha que o conjunto de dados anônimo a ser disponibilizado é construído iterativamente por meio da aplicação de operações de especialização. Cada especialização divide um valor geral em diferentes filhos, de modo que há algum ganho de informação $IG(s)$, e ao mesmo tempo perda de privacidade $PL(s)$. A métrica $IGPL(s)$ [22] busca justamente encontrar a melhor especialização s que maximiza o ganho de informação para cada perda de privacidade. Dessa forma ela pode ser definida como:

$$IGPL(s) = \frac{IG(s)}{PL(s)+1}$$

A escolha de $IG(s)$ e $PL(s)$ depende da métrica de informação e do modelo de privacidade a serem adotados.

3.6. Modelos de Privacidade Sintáticos

Modelos de privacidade sintáticos têm como objetivo estabelecer uma determinada condição, a qual os dados devem satisfazer, após um processo de anonimização. Tais modelos utilizam, na maioria das vezes, generalização e/ou supressão nos dados até uma condição sintática ser atendida, de modo que o conhecimento adversário se torna restrito na descoberta de atributos sensíveis a partir de semi-identificadores.

Nesta seção iremos apresentar alguns desses modelos de privacidade sintáticos e detalhar seus pontos positivos e negativos com o auxílio de exemplos. Vale ressaltar que serão vistos apenas modelos de como preservar a privacidade e não algoritmos específicos de como atingir a propriedade específica de cada um deles.

3.6.1. k-anonimato

O modelo de privacidade k-anonimato é o mais conhecido no campo da anonimização de dados. Foi proposto por Sweene [42] como forma de proteção ao ataque de ligação ao registro. Esse modelo assegura que, para cada combinação de valores de semi-identificadores, existem pelo menos k registros no conjunto de dados publicado, formando uma classe de equivalência. O k-anonimato atua sobre o princípio da indistinguibilidade, isto é, cada registro em um conjunto de dados k-anônimo é indistinguível de pelo menos outros k-1 registros em relação ao conjunto de semi-identificadores. Dessa forma, garante-se que cada registro não pode ser ligado a um indivíduo por um adversário com probabilidade maior que $\frac{1}{k}$.

Neste modelo, o valor de k define o nível de privacidade e, ao mesmo tempo, atua diretamente sobre a perda de informação. Assim, quanto maior o valor de k, maior será a privacidade e menor a utilidade dos dados. Não existem abordagens analíticas para determinar o melhor valor de k [13]. A complexidade na escolha desse valor depende de muitos critérios, como por exemplo dos requisitos de privacidade provenientes do detentor dos dados, dos requisitos de utilidade por parte de analistas, testadores, pesquisadores, etc., do nível de generalização, dentre outros critérios.

Para ilustrar a utilização desse modelo, iremos considerar que se deseja publicar os dados da Tabela 3.4 seguindo o modelo 2-anonimato. Vamos considerar que os identificadores explícitos são Placa, Motorista e CPF e como atributos sensíveis Tipo de Multa e Valor da Multa e, consequentemente, os demais como atributos semi-identificadores: Data de Nascimento e Data da Infração.

	Placa	**Motorista**	**CPF**	**Data de Nascimento**	**Data da Infração**	**Tipo de Multa**	**Valor da Multa (R$)**
1	HXR-1542	José Pereira	258.568.856	14/03/1977	03/01/2013	1	170
2	HTS-5864	Jorge Cury	566.548.584	04/03/1977	03/01/2013	2	250
3	HUI-5846	Paula Maria	384.987.687	24/05/1977	03/01/2013	1	170
4	HTR-5874	Jandira Lima	054.864.576	20/04/1978	04/01/2013	1	170
5	HOI-6845	José Sá	244.684.876	22/05/1978	04/01/2013	2	250
6	HQO-5846	Kilvia Mota	276.684.159	13/05/1978	05/01/2013	2	250
7	HUY-8545	José Pereira	538.687.045	15/05/1978	05/01/2013	1	170

Tabela 3.4. Dados sobre infrações de trânsito (Fonte: [11]).

Após aplicar supressão nos identificadores explícitos e generalização nos atributos sensíveis a Tabela 3.5 é gerada. Nesta tabela podemos perceber quatro classes de equivalência para os semi-identificadores: Classe A = "$03/1977,01/2013$" nas linhas 1 e 2; Classe B = "$05/1977,01/2013$" registro 3; Classe C = "$04/1978,01/2013$" com o registro 4 e Classe D = "$05/1978,01/2013$" nas linhas 5, 6 e 7. Perceba que, mesmo após aplicar algum nível de generalização, a regra 2-anonimato está sendo violada para os registros das linhas 3 e 4. Se algum atacante tiver disponível como conhecimento prévio a data de nascimento e souber que esses indivíduos estão representados nos dados publicados, apesar da generalização, ele consegue inferir o registro referente aos indivíduos. Neste caso, alguma outra operação de anonimização complementar deve ser adotada para esses grupos, como a supressão dos registro conforme a Tabela 3.6 que atende ao critério $k = 2$.

É mostrado em [20] que este modelo é efetivo contra ataques de ligação ao registro, porém não é um modelo adequado para prevenir ataques de ligação ao atributo. Considere, por exemplo, a Tabela 3.6 que atende ao modelo 2-anonimato e cujo adversário tenha conhecimento prévio que José Sá (linha 5) nasceu em 1978 e que foi multado em Janeiro de 2013. O adversário consegue inferir que o valor da multa de José foi de R$ 250 com probabilidade $\frac{2}{3}$, i.e., maior do que $\frac{1}{2}$ exigido pelo modelo. Outro fator que não é considerado pelo modelo é o caso de uma publicação possuir mais de um registro referente ao mesmo indivíduo. Destaca-se, ainda, que o problema de encontrar uma k-anonimização ótima é considerado NP-difícil, conforme demonstrado em [36].

Como forma de contornar algumas das desvantagens do modelo k-anonimato, fo-

	Placa	Motorista	CPF	Data de Nascimento	Data da Infração	Tipo de Multa	Valor da Multa (R$)
1	*	*	*	03/1977	01/2013	1	170
2	*	*	*	03/1977	01/2013	2	250
3	*	*	*	05/1977	01/2013	1	170
4	*	*	*	04/1978	01/2013	1	170
5	*	*	*	05/1978	01/2013	2	250
6	*	*	*	05/1978	01/2013	2	250
7	*	*	*	05/1978	01/2013	1	170

Tabela 3.5. Dados sobre informações de trânsito anonimizados (Fonte: [11]).

	Placa	Motorista	CPF	Data de Nascimento	Data da Infração	Tipo de Multa	Valor da Multa (R$)
1	*	*	*	03/1977	01/2013	1	170
2	*	*	*	03/1977	01/2013	2	250
3	*	*	*	*	*	*	*
4	*	*	*	*	*	*	*
5	*	*	*	05/1978	01/2013	2	250
6	*	*	*	05/1978	01/2013	2	250
7	*	*	*	05/1978	01/2013	1	170

Tabela 3.6. Tabela no modelo *2-anonimato* (Fonte: [11]).

ram propostos outros modelos de privacidade sintáticos, como l-diversidade, t-proximidade, δ-presença, entre outros.

3.6.2. l-diversidade

O modelo l-diversidade [34] busca prover proteção contra ataques de ligação ao atributo, i.e., os casos em que um adversário pode inferir informações sensíveis sobre registros mesmo sem identificá-los. Surgiu como forma de sanar essa limitação do k-anonimato. Para evitar esse tipo de descoberta, o modelo exige que cada classe de equivalência possua, pelo menos, l valores distintos para cada atributo sensível. Isto garante que um atacante, mesmo com conhecimento prévio que lhe permita descobrir a classe de equivalência de um indivíduo, não consiga inferir o atributo sensível do mesmo com probabilidade maior que $\frac{1}{l}$.

Na Tabela 3.7, onde os atributos Idade, CEP e Cidade foram classificados como semi-identificadores e o atributo Doença como sensível, os registros estão anonimizados segundo o modelo 4-anonimato. Porém, perceba que se um adversário possuir conhecimento prévio de que o CEP de um dado indivíduo é 540020, é possível deduzir que este pertence à classe de equivalência "$40,540020$" - registros das linhas 5 a 8. E, com isso, ele consegue inferir que o indivíduo sofre de bronquite.

Ao converter a Tabela 3.7 para o modelo 3-diversidade, não é preciso fazer modificações para a classe de equivalência "$85,560001$", pois esta já possui quatro valores distintos para o atributo sensível, satisfazendo a condição imposta pelo modelo. Já para a

classe "40, 540020" é necessário aplicar alguma técnica para satisfazer o requisito. Uma possibilidade seria suprimir os registros das linhas de 5 a 8. Uma outra solução é demonstrada na Tabela 3.8 onde os atributos sensíveis das linhas 5 e 6 foram alterados de forma que, agora, satisfazem ao modelo.

	Idade	**CEP**	**Cidade**	**Doença**
1	<85	560001	*	Sinusite
2	<85	560001	*	Gripe
3	<85	560001	*	Diabetes
4	<85	560001	*	Hérnia
5	<40	540020	*	Bronquite
6	<40	540020	*	Bronquite
7	<40	540020	*	Bronquite
8	<40	540020	*	Bronquite

Tabela 3.7. Conjunto de dados no modelo 4-anonimato (Fonte: [44]).

	Idade	**CEP**	**Cidade**	**Doença**
1	<85	560001	*	Sinusite
2	<85	560001	*	Gripe
3	<85	560001	*	Diabetes
4	<85	560001	*	Hérnia
5	<40	540020	*	Sinusite
6	<40	540020	*	Diabetes
7	<40	540020	*	Bronquite
8	<40	540020	*	Bronquite

Tabela 3.8. Conjunto de dados no modelo 4-anonimato e 3-diversidade (Fonte: [44]).

Como mencionado em [44] e [20], o l-diversidade apresenta alguns pontos não cobertos pelo modelo:

- Impacto na utilidade dos dados quando o modelo é aplicado em cenários onde há grande número de repetições para o valor de um atributo sensível e pouco/nenhum de outros valores, pois neste caso é necessário introduzir grande distorção ou supressão nos dados.
- *Skewness Attack*: ocorre quando o atacante tem conhecimento prévio e descobre tanto a classe de equivalência de um indivíduo quanto a distribuição dos atributos sensíveis, que pode ser obtida apenas analisando a tabela publicada. De posse dessas duas informações, o atacante pode obter o valor de um atributo sensível com uma chance maior do que a proporcionada pela distribuição global. Seja, por exemplo, para o modelo 2-diversidade uma tabela onde apenas dois por cento dos indivíduos tem diagnóstico positivo para o atributo sensível HIV. Se uma determinada classe de equivalência com quatro registros apresenta dois destes com diagnóstico positivo e dois com negativo, infere-se com uma chance de $\frac{1}{2}$ que o indivíduo é positivo contra a probabilidade global de $\frac{1}{50}$.

- Ataque de similaridade: ocorre quando, mesmo que os atributos sensíveis sejam distintos, qualquer um deles fornece uma informação sensível ao atacante, e.g., no caso em que o atributo sensível é uma doença no modelo 2-diversidade e os valores para a classe de equivalência de um dado indivíduo descoberta pelo atacante a partir de conhecimento prévio são úlcera ou gastrite, por similaridade ele consegue inferir que o indivíduo tem uma doença de estômago.
- Incapacidade de lidar com a semântica de relação dos novos valores ao substituir os originais, e.g., na Tabela 3.8 onde os valores Bronquite (linhas 5 e 6) foram substituídos por Sinusite e Diabetes.

3.6.3. t-proximidade

O modelo t-proximidade [33] propõe-se a corrigir algumas limitações do l-diversidade no que diz respeito à proteção contra *skewness attack*, onde o adversário pode inferir informações sobre atributos sensíveis a partir do conhecimento da frequência de ocorrência dos atributos na tabela (conforme exemplificado na Seção 3.6.2). Para isso, esse modelo visa assegurar que a distribuição dos dados de um atributo sensível em cada classe de equivalência seja próxima à sua distribuição global (i.e. na tabela completa).

A distância máxima entre as classes e a distribuição global é definida através do parâmetro t. Para mensurar a diferença entre a distribuição da classe de equivalência e a global é utilizada a "distância global de retaguarda" (*EMD: Earth Mover Distance*) cujo resultado contém valores reais no intervalo $[0,1]$, observando que, quanto maior o valor da distância, mais fraca é a proteção. No trabalho [33] EMD é adaptada a calcular a distância $D[P,Q]$ onde $P = \{p_1, p_2, ..., p_m\}$ é a distribuição dos atributos de uma dada classe, m o número de atributos sensíveis e $Q = \{q_1, q_2, ..., q_m\}$) a distribuição global.

As principais limitações [20] do t-proximidade são resumidas em: (i) falta de flexibilidade para especificação de diferentes níveis de privacidade para cada atributo sensível, (ii) a função EMD não é adequada para ataques de ligação ao atributo quando estes são numéricos e (iii) garantir o t-proximidade poderá degradar bastante a utilidade para garantir a mesma distribuição em todos as classes de equivalência.

3.6.4. δ-presença

Este modelo foi proposto em [38] e objetiva evitar a vinculação de um indivíduo a uma tabela publicada, ou seja, busca proteger a privacidade de indivíduos contra o ataque de ligação à tabela. O modelo δ-presença define o limite $\delta = (\delta_{max}, \delta_{min})$ para a probabilidade de um adversário inferir a presença de um indivíduo em um conjunto de dados tabulados. Este modelo pode prevenir indiretamente ataques de ligações ao registro e ao atributo, visto que um atacante possui no máximo $\delta\%$ de confiança de que o registro da vítima está presente na tabela publicada, então a probabilidade de uma ligação ser bem-sucedida ao registro e ao atributo sensível é no máximo $\delta\%$.

Para exemplificar um ataque de ligação à tabela, vamos utilizar um exemplo de [20]. Suponha que o detentor dos dados tenha liberado informações anonimizadas de pacientes na Tabela T (Tabela 3.9) e, ainda, que um adversário tenha acesso a dados públicos da Tabela E (Tabela 3.10) e sabe que todos os registros da primeira estão contidos na segunda ($T \subset E$). Com isso consegue-se deduzir que a probabilidade de Alice estar

presente em T é $\frac{4}{5} = 0.8$ pois há 4 registros em T e 5 registros em E contendo a classe de Alice: "$Artista, Feminino, [30-35)$". Pelo mesmo raciocínio, a probabilidade de Bob estar presente é de $\frac{3}{4} = 0.75$.

Profissão	Gênero	Idade	Doença
Profissional	Masculino	[35-40)	Hepatite
Profissional	Masculino	[35-40)	Hepatite
Profissional	Masculino	[35-40)	HIV
Artista	Feminino	[30-35)	Gripe
Artista	Feminino	[30-35)	HIV
Artista	Feminino	[30-35)	HIV
Artista	Feminino	[30-35)	HIV

Tabela 3.9. Tabela de pacientes no formato 3-anonimimato (Fonte: [20]).

Nome	Profissão	Gênero	Idade
Alice	Artista	Feminino	[30-35)
Bob	Professional	Masculino	[35-40)
Cathy	Artista	Feminino	[30-35)
Doug	Professional	Masculino	[35-40)
Emily	Artista	Feminino	[30-35)
Fred	Professional	Masculino	[35-40)
Gladys	Artista	Feminino	[30-35)
Henry	Professional	Masculino	[35-40)
Irene	Artista	Feminino	[30-35)

Tabela 3.10. Tabela externa no formato 4-anonimato (Fonte: [20]).

3.7. Modelo de Privacidade Diferencial

Os modelos de privacidade apresentados até aqui consideram que a violação de privacidade ocorre quando dados são publicados em formatos tabulados e o adversário utiliza informações de fontes externas para reidentificar indivíduos. Sob outra perspectiva, a Privacidade Diferencial investiga a ideia de publicar resultados de consultas, ao invés de dados tabulados, de tal forma que são adicionados ruídos a esses resultados. Em outras palavras, os dados provenientes de consultas são perturbados como forma de garantir a privacidade dos indivíduos. Assim, um atacante não será capaz de concluir algo com 100% de confiança. A sua principal convicção é de que as conclusões obtidas sobre um indivíduo são referentes aos dados de toda a tabela, e não apenas a um registro em particular. Por esse motivo, o modelo de privacidade em questão propõe evitar ataques probabilísticos.

3.7.1. Conceitos Básicos

A Privacidade Diferencial é um modelo matemático proposto em [17] que fornece sólidas garantias de privacidade. O objetivo desse modelo é disponibilizar informações estatísticas sobre um conjunto de dados sem comprometer a privacidade de seus indivíduos.

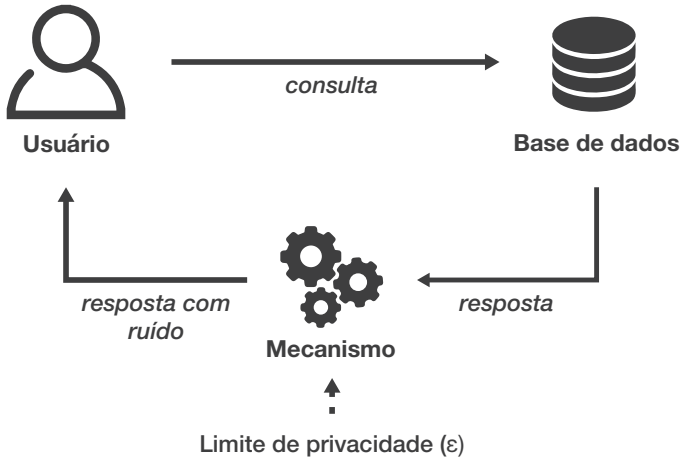

Figura 3.7. Ambiente interativo no modelo de Privacidade Diferencial.

A Privacidade Diferencial é satisfeita por um algoritmo aleatório, geralmente chamado de mecanismo. Este modelo foi projetado em um ambiente interativo, onde os usuários submetem consultas a um conjunto de dados e este, por sua vez, responde por meio de um mecanismo de anonimização. Este ambiente interativo é apresentado na Figura 3.7. Esse mecanismo proporcionará a privacidade, introduzindo "aleatoriedade" e protegendo os resultados das consultas sobre o conjunto de dados original.

A Privacidade Diferencial assegura que qualquer sequência de resultados (isto é, resposta de consultas) é igualmente possível de acontecer independente da presença de qualquer indivíduo no conjunto de dados [19]. Assim, a adição ou remoção de um indivíduo não afetará consideravelmente o resultado de qualquer análise estatística realizada no conjunto de dados [14]. Portanto, um adversário não deve ser capaz de aprender nada sobre um indivíduo específico que ele já não poderia ter aprendido antes sem acesso ao conjunto de dados.

3.7.2. Definição Formal

Dado um algoritmo aleatório (mecanismo) M, este mecanismo garante ε-Privacidade Diferencial se para todos os conjuntos de dados vizinhos D_1 e D_2 no conjunto de dados, estes diferem de no máximo um elemento, e para todo S contido na variação de resultados de M, isto é, para todo S $\subseteq$ Range(M),

$$Pr[M(D_1) \in S] \leq exp(\varepsilon) \times Pr[M(D_2) \in S],$$

onde Pr é a probabilidade dada a partir da "aleatoriedade" de M. Em resumo, a definição formal afirma que a diferença entre as probabilidades de uma consulta retornar o mesmo resultado em dois conjuntos de dados é limitada pelo parâmetro ε. Dessa forma, para qualquer par de entradas que diferem de apenas um registro, para cada saída, um adversário não poderá distinguir entre os conjuntos de dados D_1 e D_2 baseado apenas na resposta fornecida pelo mecanismo.

A Figura 3.8 mostra um exemplo de probabilidades de saída de um algoritmo M, nos conjuntos de dados vizinhos D_1 e D_2, a partir de um valor de ε. O algoritmo M fornece

garantias de privacidade adicionando ruído aleatório no seu retorno, i.e., $M(D) = f(D) + ruido$, onde f é a resposta de uma consulta realizada por um usuário. Nesse exemplo, as probabilidades seguem o mecanismo de Laplace, o que resulta em uma distribuição com pico mais acentuado do que uma distribuição normal. O conceito de mecanismo é definido na Seção 3.7.3.

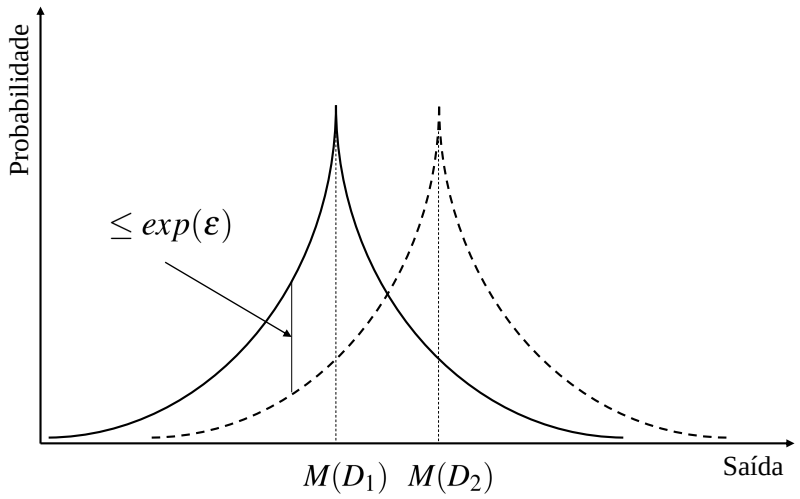

Figura 3.8. Probabilidades de saída de um algoritmo aleatório M sobre os conjuntos de dados vizinhos D_1 e D_2.

Se um indivíduo, portanto, escolhe participar de um conjunto de dados D onde análises estatísticas serão feitas através de um mecanismo que é ε-Diferencialmente Privado, esse mecanismo irá garantir que não haverá um aumento na probabilidade de violação de privacidade se comparado com a probabilidade quando o indivíduo escolhesse não participar do conjunto de dados. Dessa forma, podemos concluir que, como a Privacidade Diferencial é uma propriedade estatística sobre como o mecanismo funciona, as garantias que ela oferece são altas, inclusive essas garantias não dependem de poder computacional ou informações que um atacante possa ter obtido.

A definição formal de Privacidade Diferencial mostrada acima não leva em consideração como podemos escolher o parâmetro ε, principalmente porque esse parâmetro não possui correlação explícita com a privacidade dos indivíduos como em outras técnicas vistas. Esse parâmetro depende da consulta que está sendo feita e dos próprios dados que estão no conjunto de dados. A literatura em geral concorda que o valor de ε deva ser pequeno, como por exemplo 0.01, 0.1 ou até logarítmo natural $ln2$ ou $ln3$ [18]. Quanto menor o valor de ε, maior a privacidade. Dessa forma a escolha do valor de ε deve ser experimental e algumas vezes este valor é encontrado empiricamente, portanto, para cada mecanismo, deve ser feita uma análise para escolher o parâmetro adequado utilizando métricas [39] para avaliar a precisão da resposta do mecanismo com diversos valores de ε [30].

3.7.3. Mecanismo e Sensibilidade

Como foi dito anteriormente, a Privacidade Diferencial foi definida em um modelo interativo, onde o usuário submete consultas a uma base de dados D e, consequentemente, um

determinado mecanismo fornece uma resposta ε-Diferencialmente Privada. Porém, existem diversas formas de se atingir a Privacidade Diferencial através de um mecanismo. O objetivo das técnicas que utilizam esse modelo de privacidade é criar um mecanismo M que irá adicionar um ruído adequado para produzir uma resposta a uma consulta f feita pelo indivíduo, de forma que esse ruído seja independente do conjunto de dados D.

A quantidade de ruído necessária depende do tipo de consulta f aplicada sobre um conjunto de dados. Dessa forma precisamos definir o que é a sensibilidade de um conjunto de dados D. Antes disso, porém, precisamos entender o que são conjuntos de dados vizinhos.

Definição 2 *Dado um conjunto de dados D, todos os conjuntos de dados D_i decorrentes da remoção de um indivíduo i do conjunto de dados original D são definidos como vizinhos.*

Por exemplo, considere o conjunto de dados D na Figura 3.9a. Um de seus vizinhos pode ser obtido pela remoção do registro de ID $= 3$, resultando na Figura 3.9b, uma vez que não houve alteração nos valores dos registros.

ID	Peso (Kg)	Altura (m)
1	87,2	1,70
2	81,2	1,62
3	74,2	1,75
4	60,0	1,61
5	78,5	1,58

(a)

ID	Peso (Kg)	Altura (m)
1	87,2	1,70
2	81,2	1,62
4	60,0	1,61
5	78,5	1,58

(b)

Figura 3.9. Exemplo de conjuntos de dados vizinhos.

Definição 3 *Seja D o domínio de todos os conjuntos de dados. Seja f uma função de consulta que mapeia conjuntos de dados a vetores de números reais. A sensibilidade global da função f é:*

$$\Delta f = \max_{x,y \in D} \| f(x) - f(y) \|_1$$

para todo x, y diferindo de no máximo um elemento, ou seja, vizinhos. [18].

A sensibilidade então vai medir quanta diferença um usuário faz ao ser removido do conjunto de dados na resposta da função de consulta. Isso é fundamental para o cálculo adequado do ruído a ser adicionado pelo mecanismo, uma vez que quanto maior o valor de Δf, mais ruído terá de ser adicionado à resposta do mecanismo para mascarar a remoção de um indivíduo, de forma a assegurar a privacidade do mesmo [15].

O mecanismo de Laplace é o método mais comum e simples para alcançar a Privacidade Diferencial. A adição de ruído é baseada na geração de uma variável aleatória da distribuição de Laplace com média μ e escala b de forma que

$$Laplace_{\mu,b}(x) = \frac{1}{2b}\exp(-\frac{|x|}{b})$$

Podemos então definir formalmente o mecanismo de Laplace.

Definição 4 *Dada uma função de consulta $f : D \to \Re$, o mecanismo de Laplace M:*

$$M_f(D) = f(D) + Laplace(0, \Delta f/\varepsilon)$$

fornece ε-Privacidade Diferencial. Onde $Laplace(0, \Delta f/\varepsilon)$ retorna uma variável aleatória da distribuição de Laplace com média zero e escala $\Delta f/\varepsilon$.

3.7.4. Exemplo Explicativo

Considere o seguinte exemplo explicativo de utilização da Privacidade Diferencial em pequena escala. Suponha um conjunto de dados da Receita Federal, que contém o número de imóveis que um determinado indivíduo declarou, conforme Figura 3.10.

ID	Nome	Nº Imóveis
1	Roney	4
2	André	2
3	Leo	7
4	Bruno	1

Figura 3.10. Exemplo de conjunto de dados original contendo o número de imóveis de cada indivíduo.

Suponha também que a consulta f a ser realizada sobre essa base de dados retorna a soma de todos os imóveis de todos os indivíduos. Para aplicar a Privacidade Diferencial sobre esses dados é necessário calcular f para cada vizinho do conjunto original. A resposta real da consulta é 14. A Figura 3.11 mostra os conjuntos de dados vizinhos gerados a partir da base original e suas respectivas respostas da consulta f.

Para garantir a privacidade acerca de um indivíduo, ele participando ou não do conjunto de dados, é preciso calcular a variação máxima que a ausência do indivíduo provoca no resultado da consulta. Essa variação é proveniente da remoção do registro de $ID = 3$. Em seguida, é necessário calcular a sensibilidade da consulta aplicada ao conjunto de dados. Conforme definida nesta seção, a sensibilidade é calculada pela maior diferença $|f(D) - f(D_i)|$, e ocorre quando $i = 3$, pois gera como resultado $|14 - 7| = 7$. Por fim, o ruído a ser adicionado para atender ao modelo de Privacidade Diferencial, utilizando um mecanismo de Laplace, deve ser igual a $Laplace(0, \frac{7}{\varepsilon})$.

O parâmetro ε é definido pelo detentor dos dados. A Tabela 3.11 apresenta cinco exemplos de ruído, respostas e probabilidade de ocorrência após a aplicação da Privacidade Diferencial sobre o conjunto de dados original da Figura 3.10, considerando $\varepsilon = 1$.

ID	Nome	Nº Imóveis
2	André	2
3	Leo	7
4	Bruno	1

$f(D_1) = 2 + 7 + 1 = 10$

ID	Nome	Nº Imóveis
1	Roney	4
3	Leo	7
4	Bruno	1

$f(D_2) = 4 + 7 + 1 = 12$

ID	Nome	Nº Imóveis
1	Roney	4
2	André	2
4	Bruno	1

$f(D_3) = 4 + 2 + 1 = 7$

ID	Nome	Nº Imóveis
1	Roney	4
2	André	2
3	Leo	7

$f(D_4) = 4 + 2 + 7 = 13$

Figura 3.11. Conjuntos de dados vizinhos gerados a partir da base original e suas respectivas respostas da consulta f (soma).

Ruído	*f(D) + ruído*	*Pr(f(D) + ruído)%*
-4,58	9,42	3,70
-0,15	13,85	6,98
12,15	26,15	1,25
-6,43	7,57	2,85
2,89	16,89	4,72

Tabela 3.11. Cinco possíveis valores de ruído, resposta e probabilidade de ocorrência após a aplicação da Privacidade Diferencial.

Assim, após utilizar o mecanismo de Laplace, o valor de ruído de $-4,58$ possui probabilidade de ocorrência de $3,7\%$ sobre o valor original da consulta f (cuja soma do número de imóveis original é igual a 14), resultando em um valor anonimizado de $9,42$ imóveis. De forma análoga, o valor de ruído de $-0,15$ possui uma probabilidade um pouco maior de ocorrência $(6,98\%)$, caso a mesma consulta seja realizada nesse conjunto de dados, conforme mostra a Tabela 3.11.

3.7.5. Limitações e Desafios

A principal promessa da Privacidade Diferencial, como foi dito anteriormente, é que, se podemos extrair informações sem os dados de um indivíduo, então a privacidade do mesmo não foi violada. Contudo, nada impede que um adversário com conhecimento externo possa descobrir algo sobre um indivíduo sem ferir essa promessa da Privacidade Diferencial, exatamente porque esse indivíduo poderia nem fazer parte do conjunto de dados, mas poderia ter semelhanças com os outros indivíduos e o resultado do mecanismo ser utilizado para descobrir informações. Esse é um ponto de bastante criticidade da abordagem, porém isso não é considerado uma violação, já que o modelo de privacidade é relativo, i.e., participar ou não de um conjunto de dados que será submetido para análises somente aumenta ligeiramente o risco de descoberta [12].

No entanto existem alguns problemas na utilização da Privacidade Diferencial. O principal deles é o cálculo da sensibilidade da consulta que pode ser complexo e gerar muito ruído, o que irá afetar significativamente a utilidade dos dados. Apesar de existirem técnicas para minimizar essa perda, ainda é difícil balancear o ruído adicionado e manter a garantia da privacidade. O valor do parâmetro ε na prática também é difícil de ser estimado, especialmente porque existem poucos trabalhos relacionados e também pelo fato de que o seu valor não é uma medida direta de privacidade e sim um limitante do impacto que um usuário faz em um conjunto de dados. Os valores retornados pelo mecanismo podem não ser adequados para utilização em áreas específicas devido à natureza (probabilística) incerta de como esses dados são gerados. Então muitas vezes, nas aplicações reais, as técnicas de anonimização como *k*-anonimato ainda são bastante utilizadas.

3.8. Aplicações

Existem implementações práticas de aplicações que visam garantir a privacidade dos usuários diante de um mundo cada vez mais conectado e que cada vez mais gera informações através dos navegadores de internet e *smartphones*. Um exemplo desse tipo de aplicação é o *ARX*[1], um software livre que tem por objetivo prover fácil compreensão e suporte a diversos modelos de privacidade. O *ARX* dá suporte tanto para modelos de privacidade sintáticos, isto é, *k*-anonimato, *l*-diversidade, *t*-proximidade e δ-presença, como para modelos de privacidade semânticos, i.e ε-Privacidade Diferencial. O *ARX* também possibilita a anonimização dos dados após aplicadas operações de generalização e supressão, conforme Figuras 3.12 e 3.13. Nesse caso foi utilizado o modelo de privacidade *k*-anonimato.

	sex	age	race	marital-status	education	native-coun...	workclass
74	Male	52	White	Married-civ-spouse	Doctorate	United-States	Local-gov
75	Male	52	White	Married-civ-spouse	Masters	United-States	Local-gov
76	Male	51	White	Married-civ-spouse	Some-college	United-States	State-gov
77	Male	51	White	Married-civ-spouse	Bachelors	United-States	State-gov
78	Male	51	White	Married-civ-spouse	Bachelors	United-States	Local-gov
79	Male	51	White	Married-civ-spouse	Bachelors	United-States	Local-gov

Figura 3.12. Exemplo de dados originais no ARX.

	sex	age	race	marital-status	education	native-coun...	workclass
74	Male	[51, 60]	White	Spouse present	Higher education	North America	Government
75	Male	[51, 60]	White	Spouse present	Higher education	North America	Government
76	Male	[51, 60]	White	Spouse present	Higher education	North America	Government
77	Male	[51, 60]	White	Spouse present	Higher education	North America	Government
78	Male	[51, 60]	White	Spouse present	Higher education	North America	Government
79	Male	[51, 60]	White	Spouse present	Higher education	North America	Government

Figura 3.13. Exemplo de dados anonimizados pelo ARX.

O ARX também dispõe de métodos para a análise da utilidade desses dados e análise dos riscos de identificação (Figuras 3.14 e 3.15). Os respectivos resultados mostram a quantidade de registros em risco de violação de privacidade, o maior risco de violação e a porcentagem de sucesso caso um atacante deseje associar informações contidas na mesma

[1] http://arx.deidentifier.org/

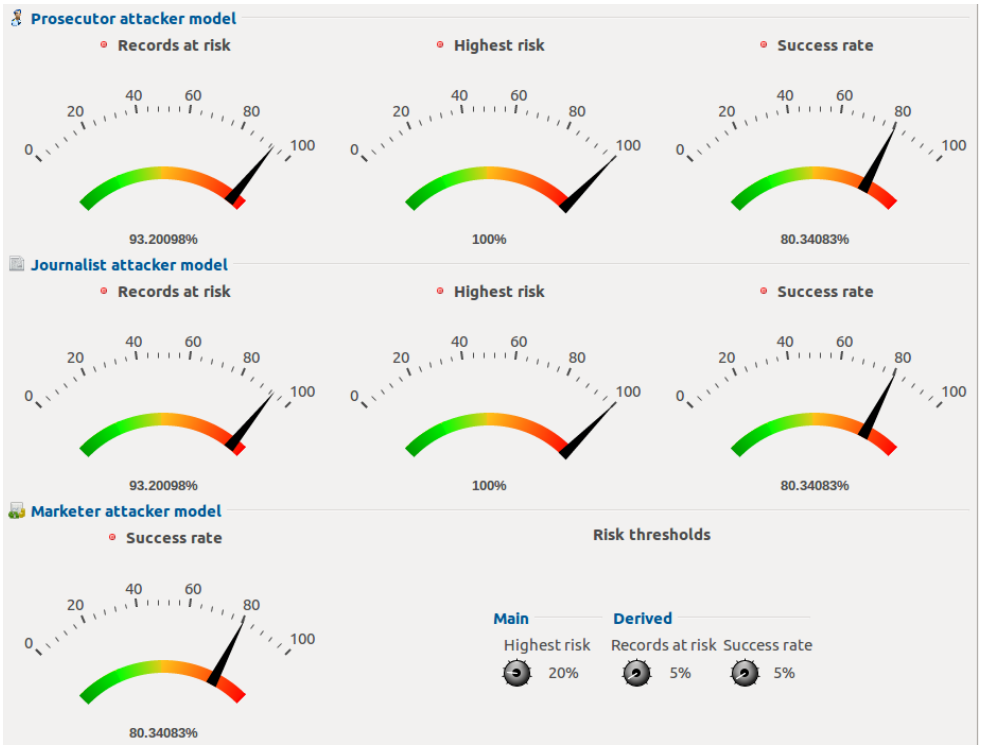

Figura 3.14. Exemplo de análise de riscos nos dados originais feita pelo ARX.

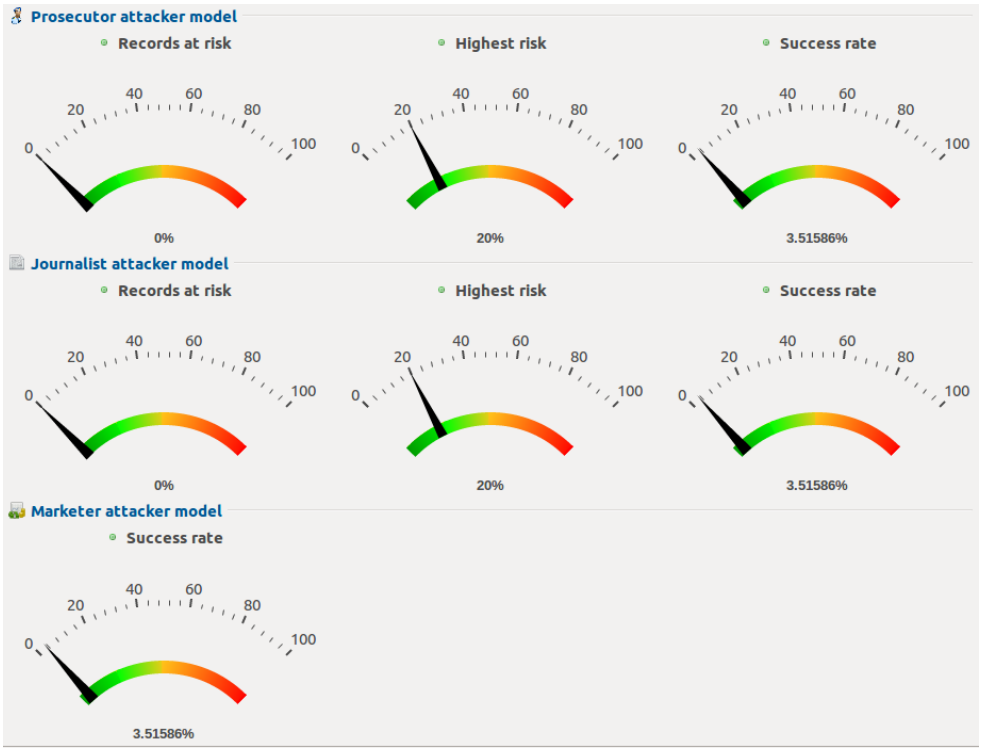

Figura 3.15. Exemplo de análise de riscos nos dados anonimizados feita pelo ARX.

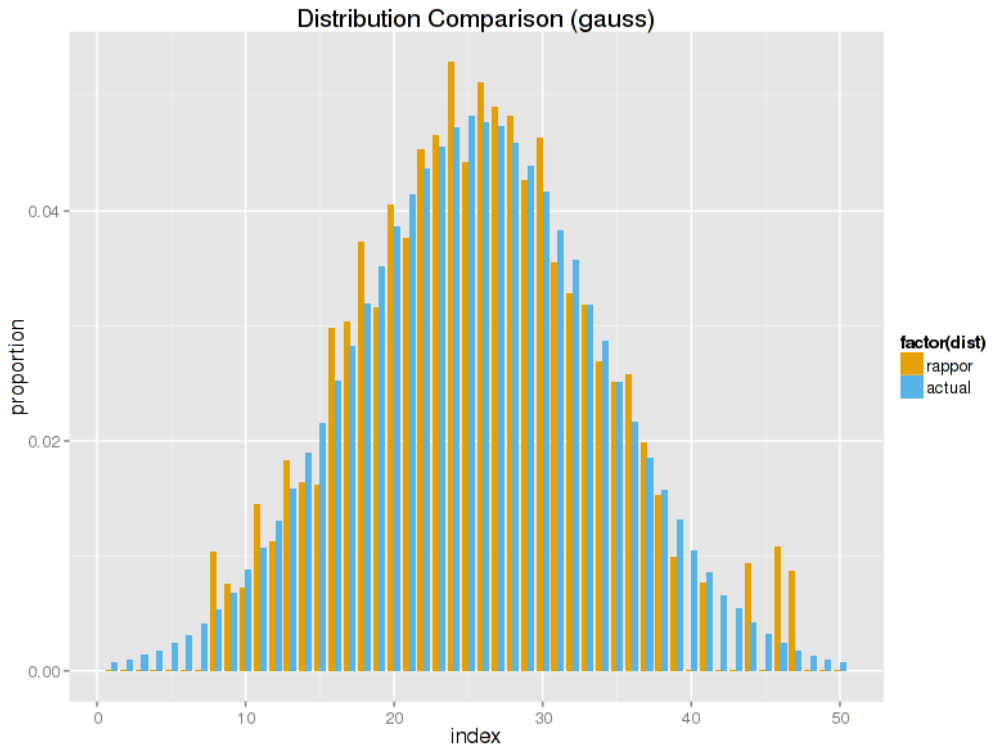

Figura 3.16. Exemplo de análise de distribuição feita pelo RAPPOR.

tabela. Além do Software com interface gráfica, o *ARX* também disponibiliza uma API que é capaz de possibilitar a anonimização de dados para qualquer aplicação Java.

Por outro lado, muitas organizações detentoras de dados têm desenvolvido aplicações para garantir a privacidade de dados coletados de seus clientes utilizando o modelo de Privacidade Diferencial. A Microsoft, por exemplo, desenvolveu o *PINQ* [35], uma plataforma de análise de dados projetada para fornecer garantias de Privacidade Diferencial. O *PINQ* utiliza um sistema de consultas próprio, semelhante ao SQL, denominado LINQ, e age como uma camada intermediária entre uma base de dados e o detentor dos dados, permitindo-o realizar consultas sem comprometer a privacidade dos indivíduos pertencentes a base. Um protótipo do *PINQ* está disponível para download no site oficial da Microsoft para experimentação. O *PINQ* possui vários exemplos de aplicações, além de ter sido projetado para que possa ser utilizado mesmo sem conhecimentos profundos sobre privacidade.

Com a grande necessidade de monitorar estatísticas de usuários, por exemplo configurações do navegador, o Google também lançou uma ferramenta denominada *Rappor* (*Randomized Aggregatable Privacy Preserving Ordinal Responses*), que visa proteger essas informações e facilitar a vida de quem precisa de dados coletados através de aplicações. Essa ferramenta utiliza a perturbação na coleta de dados, mantendo as informações estatísticas necessárias para realizar suas análises e preservando a privacidade dos usuários que utilizam seu navegador. É possível encontrar a implementação do demo no Github, que faz uma simulação e uma análise dos dados utilizando Python e R. Um exem-

plo de uma das análises feita pelo *Rappor* é vista na Figura 3.16, onde pode-se visualizar a distribuição dos dados originais, gerados pela simulação, e a distribuição dos dados alterados pelo *Rappor* para garantir a privacidade diferencial.

Outros dois exemplos recentes de aplicações no mundo real que também utilizam as técnicas vistas neste capítulo são apresentados em [37] e [10]. No primeiro, denominado $DP-WHERE$, os autores buscam preservar a privacidade de dados de mobilidade utilizando a rede de dados de celulares. Eles demonstram que é possível balancear privacidade e utilidade em aplicações práticas utilizando grandes volumes de dados. O segundo trabalho propõe liberar estatísticas acerca de 70 milhões de senhas utilizando Privacidade Diferencial. Os autores provam que o mecanismo proposto introduz mínima distorção nos dados, assegurando que a lista de senhas disponibilizada é muito próxima da lista original.

Por fim, a Apple anunciou em 2016 que vem aplicando Privacidade Diferencial na coleta de dados dos usuários do iOS 10. Ela foi a primeira a adotar o modelo de Privacidade Diferencial em larga escala, apesar da Microsoft já estudar o assunto há um certo tempo. Dessa forma, recursos como *Siri* e até o *QuickType* poderão predizer melhor as palavras que, por exemplo, um determinado conjunto de usuários mais utiliza. Com a utilização de Privacidade Diferencial, a Apple está enviando mais informações dos dispositivos para seus servidores que antes, mas garantindo a privacidade de seus usuários através da Privacidade Diferencial.

3.9. Considerações Finais

Este capítulo conclui que a preservação de privacidade de dados acerca de indivíduos é um problema bastante desafiador. Técnicas de anonimização têm sido utilizadas para a disponibilização de dados sensíveis, buscando um balanceamento perfeito entre privacidade e utilidade, que atenda às diversas partes envolvidas no processo de disponibilização de dados. Diferentes tipos de ataques à privacidade têm sido empregados por usuários maliciosos com a intenção de violar informações sensíveis de bases de dados abertas. Para tal fim, os atacantes utilizam conhecimento que muitas vezes é imensurável, devido aos diversos cenários em que informações podem ser obtidas. Além da anonimização, foram brevemente discutidas a criptografia e a tokenização como soluções alternativas para proteção de dados sensíveis. Em abordagens de privacidade sintáticas, que estabelecem uma determinada condição a qual os dados devem pertencer antes de serem disponibilizados, os modelos de privacidade utilizam, na grande maioria dos casos, técnicas de supressão ou de generalização. Já o modelo de Privacidade Diferencial, outro paradigma apresentado, procurar fornecer soluções de preservação de privacidade de um modo mais interativo, onde uma consulta é realizada sobre conjuntos de dados e então é adicionado ruído aleatório sobre seu resultado. Aplicações no mundo real, desenvolvidas por empresas preocupadas com a privacidade de seus usuários, implementam, via de regra, o modelo de Privacidade Diferencial, principalmente devido a possibilidade de interação entre os usuários e a base de dados, o que preserva as informações estatísticas necessárias para que analistas de dados possam realizar pesquisas de maneira mais precisa sem comprometer a privacidade dos fornecedores de dados.

Finalmente, vale ressaltar que não existe a "bala de prata" que atende a qualquer

requisito de privacidade e ao mesmo tempo fornece dados úteis para qualquer tipo de análise. Também não é a mudança de paradigma que vai solucionar o problema da disponibilização de dados. Tanto o paradigma de anonimização sintática, quanto o modelo de Privacidade Diferencial apresentam questões que devem ser vistas como oportunidades de pesquisas e desenvolvimento. Não se deve abandonar uma abordagem em prol de uma outra. Avanços em ambos os paradigmas são necessários para garantir que o futuro ofereça cada vez mais proteção à privacidade de indivíduos e ao mesmo tempo haja dados úteis e disponíveis para pesquisadores, testadores, analistas de dados, e muitos outros.

Agradecimentos

Este trabalho foi parcialmente financiado com recursos da CAPES e do LSBD/UFC.

Referências

[1] 4.5 Million Records Stolen from Community Health by Chinese Hackers (2014). *Infosecurity Magazine*, https://www.infosecurity-magazine.com/news/45-million-records-stolen-from/, Acessado em: 03.04.2017.

[2] Data Protection Laws of the World. https://www.dlapiperdataprotection.com/index.html, Acessado em: 09.05.2017.

[3] Directive 95/46/EC of the European Parliament. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML, Acessado em: 09.05.2017.

[4] FIPPA Legislative Review. http://www.gov.mb.ca/chc/fippa/fippareview.html, Acessado em: 09.05.2017.

[5] HIPAA for Individuals. https://www.hhs.gov/hipaa/for-individuals/index.html, Acessado em: 09.05.2017.

[6] Most americans unwilling to give up privacy to thwart attacks (2017). *Heuters*, http://www.reuters.com/article/us-usa-cyber-poll-idUSKBN1762TQ, Acessado em: 05.04.2017.

[7] Thousands of personal details exposed in latest uk data breach blunders (2014). *Infosecurity Magazine*, https://www.infosecurity-magazine.com/news/thousands-of-personal-details, Acessado em: 03.04.2017.

[8] Aggarwal, C. C. and Yu, P. S. (2008). A framework for condensation-based anonymization of string data. *Data Min. Knowl. Discov.*, 16(3):251–275.

[9] Aggarwal, G., Feder, T., Kenthapadi, K., Motwani, R., Panigrahy, R., Thomas, D., and Zhu, A. (2005). *Anonymizing Tables*. Springer Berlin Heidelberg, Berlin, Heidelberg.

[10] Blocki, J., Datta, A., and Bonneau, J. (2016). Differentially private password frequency lists. *IACR Cryptology ePrint Archive*, 2016:153.

[11] Branco Jr, E. C., Machado, J. C., and Monteiro, J. M. (2014). Estratégias para proteção da privacidade de dados armazenados na nuvem. In *XXIX SBBD: Tópicos em Gerenciamento de Dados e Informações 2014*. Sociedade Brasileira de Computação (SBC).

[12] Clifton, C. and Tassa, T. (2013). On syntactic anonymity and differential privacy. *Transactions on Data Privacy*, 6(2):161–183.

[13] Dewri, R., Ray, I., Ray, I., and Whitley, D. (2008). On the optimal selection of k in the k-anonymity problem. In *24th ICDE International Conference on Data Engineering*, pages 1364–1366.

[14] Domingo-Ferrer, J., Sánchez, D., and Soria-Comas, J. (2016a). Database anonymization: Privacy models, data utility, and microaggregation-based inter-model connections. *Synthesis Lectures on Information Security, Privacy, & Trust*, 8(1):1–136.

[15] Domingo-Ferrer, J., Sánchez, D., and Soria-Comas, J. (2016b). Database anonymization: Privacy models, data utility, and microaggregation-based inter-model connections. *Synthesis Lectures on Information Security, Privacy, & Trust*, 8(1):1–136.

[16] Domingo-Ferrer, J. and Torra, V. (2001). A quantitative comparison of disclosure control methods for microdata. *Confidentiality, disclosure and data access: theory and practical applications for statistical agencies*, pages 111–134.

[17] Dwork, C. (2006). Differential privacy. In *33rd International Colloquium on Automata, Languages and Programming*, pages 1–12.

[18] Dwork, C. (2008). Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation*, pages 1–19. Springer.

[19] Dwork, C., Roth, A., et al. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407.

[20] Fung, B. C., Wang, K., Fu, A. W.-C., and Yu, P. S. (2010a). *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques*. Chapman & Hall/CRC, 1st edition. ISBN 978-1-4200-9148-9.

[21] Fung, B. C. M., Wang, K., Chen, R., and Yu, P. S. (2010b). Privacy-preserving data publishing: A survey of recent developments. *ACM Computer Survey*.

[22] Fung, B. C. M., Wang, K., and Yu, P. S. (2005). Top-down specialization for information and privacy preservation. In *Proceedings of the 21st International Conference on Data Engineering, ICDE 2005, 5-8 April 2005, Tokyo, Japan*, pages 205–216.

[23] Gavison, R. (1980). Privacy and the limits of law. *The Yale Law Journal*, 89(3):421–471.

[24] He, Y. and Naughton, J. F. (2009). Anonymization of set-valued data via top-down, local generalization. *Proceedings of the VLDB Endowment*, 2(1):934–945.

[25] Iyengar, V. S. (2002). Transforming data to satisfy privacy constraints. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-26, 2002, Edmonton, Alberta, Canada*, pages 279–288.

[26] Jr., R. J. B. and Agrawal, R. (2005). Data privacy through optimal k-anonymization. In *Proceedings of the 21st International Conference on Data Engineering, ICDE 2005, 5-8 April 2005, Tokyo, Japan*, pages 217–228.

[27] Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage.

[28] Larose, D. T. (2014). *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons.

[29] Laurant, C. (2003). Privacy and Human Rights 2003: an International Survey of Privacy Laws and Developments. *In Electronic Privacy Information Center*.

[30] Lee, J. and Clifton, C. (2011). *How Much Is Enough? Choosing ε for Differential Privacy*, pages 325–340. Springer Berlin Heidelberg.

[31] LeFevre, K., DeWitt, D. J., and Ramakrishnan, R. (2005). Incognito: Efficient full-domain k-anonymity. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, Baltimore, Maryland, USA, June 14-16, 2005*, pages 49–60.

[32] LeFevre, K., DeWitt, D. J., and Ramakrishnan, R. (2006). Mondrian multidimensional k-anonymity. In *Proceedings of the 22nd International Conference on Data Engineering, ICDE 2006, 3-8 April 2006, Atlanta, GA, USA*, page 25.

[33] Li, N., Li, T., and Venkatasubramanian, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. In *23th ICDE International Conference on Data Engineering (ICDE)*, pages 106–115.

[34] Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam, M. (2007). *l*-diversity: Privacy beyond *k*-anonymity. *In ACM Transactions on Knowledge Discovery from Data (TKDD)*.

[35] McSherry, F. (2010). Privacy integrated queries: an extensible platform for privacy-preserving data analysis. *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, 53(9):89–97.

[36] Meyerson, A. and Williams, R. (2004). On the complexity of optimal k-anonymity. In *Proceedings of the 23rd ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Paris, France*, pages 223–228.

[37] Mir, D. J., Isaacman, S., Cáceres, R., Martonosi, M., and Wright, R. N. (2013). DP-WHERE: differentially private modeling of human mobility. In *Proceedings of the 2013 IEEE International Conference on Big Data, 2013, Santa Clara, CA, USA*, pages 580–588.

[38] Nergiz, M. E., Atzori, M., and Clifton, C. (2007). Hiding the presence of individuals from shared databases. In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, SIGMOD '07, pages 665–676, New York, NY, USA. ACM.

[39] Nguyen, H. H., Kim, J., and Kim, Y. (2013). Differential privacy in practice. *Journal of Computing Science and Engineering*, 7(3):177–186.

[40] Solove, D. J. (2008). *Understanding Privacy*. Harvard University Press.

[41] Sweeney, L. (2002a). Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):571–588.

[42] Sweeney, L. (2002b). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, pages 557–570.

[43] Tan, V. Y. F. and Ng, S. (2007). Generic probability density function reconstruction for randomization in privacy-preserving data mining. In *Machine Learning and Data Mining in Pattern Recognition, 5th International Conference, MLDM 2007, Leipzig, Germany, July 18-20, 2007, Proceedings*, pages 76–90.

[44] Venkataramanan, N. and Shriram, A. (2016). *Data Privacy: Principles and Practice*. Chapman and Hall/CRC. ISBN 978-1-4987-2104-2.

[45] Wang, K. and Fung, B. C. M. (2006). Anonymizing sequential releases. In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20-23, 2006*, pages 414–423.

[46] Wang, K., Fung, B. C. M., and Yu, P. S. (2005). Template-based privacy preservation in classification problems. In *Proceedings of the 5th IEEE International Conference on Data Mining (ICDM 2005), 27-30 November 2005, Houston, Texas, USA*, pages 466–473.

[47] Willison, D., Emerson, C., Szala-Meneok., K., Gibson, E., Schwartz, L., and Weisbaum, K. (2008). Access to medical records for research purposes: Varying perceptions across research ethics boards. *Journal of Medical Ethics 34*, pages 308–314.

[48] Wong, R. C. and Fu, A. W. (2010). *Privacy-Preserving Data Publishing: An Overview*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers.

[49] Wong, R. C., Li, J., Fu, A. W., and Wang, K. (2006). (alpha, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing. In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA*.

[50] Xiao, X. and Tao, Y. (2006). Personalized privacy preservation. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, Chicago, Illinois, USA, June 27-29, 2006*, pages 229–240.