UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE CIÊNCIAS

DEPARTAMENTO DE COMPUTAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS DA COMPUTAÇÃO

ANDRÉ LUÍS DA COSTA MENDONÇA

UMA ABORDAGEM DE PRIVACIDADE DIFERENCIAL PARA DADOS CORRELACIONADOS UTILIZANDO TÉCNICAS DE AGRUPAMENTO

FORTALEZA

2018

ANDRÉ LUÍS DA COSTA MENDONÇA

UMA ABORDAGEM DE PRIVACIDADE DIFERENCIAL PARA DADOS CORRELACIONADOS UTILIZANDO TÉCNICAS DE AGRUPAMENTO

Dissertação apresentada ao Programa de Pós-Graduação em Ciências da Computação da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Ciência da Computação. Área de Concentração: Ciência da Computação

Orientador: Prof. Dr. Javam de Castro Machado

FORTALEZA

2018

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

M494a Mendonça, André Luís da Costa.
Uma Abordagem de Privacidade Diferencial Para Dados Correlacionados Utilizando Técnicas de Agrupamento / André Luís da Costa Mendonça. – 2018.
93 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Centro de Ciências, Programa de Pós-Graduação em Ciência da Computação, Fortaleza, 2018.
Orientação: Prof. Dr. Javam de Castro Machado.

1. Privacidade de Dados. 2. Privacidade Diferencial. 3. Dados Correlacionados. 4. Agrupamento de Dados. I. Título.

CDD 005

ANDRÉ LUÍS DA COSTA MENDONÇA

UMA ABORDAGEM DE PRIVACIDADE DIFERENCIAL PARA DADOS CORRELACIONADOS UTILIZANDO TÉCNICAS DE AGRUPAMENTO

Dissertação apresentada ao Programa de Pós-Graduação em Ciências da Computação da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Ciência da Computação. Área de Concentração: Ciência da Computação

Aprovada em: 05/03/2018

BANCA EXAMINADORA

Prof. Dr. Javam de Castro Machado (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. José Maria da Silva Monteiro Filho
Universidade Federal do Ceará (UFC)

Prof. Dr. Luciano de Andrade Barbosa
Universidade Federal de Pernambuco (UFPE)

Aos meus pais, minha namorada e a toda minha família que, com muito carinho e apoio, não mediram esforços para que eu chegasse até esta etapa de minha vida.

AGRADECIMENTOS

A Deus por ter me dado saúde e força para superar as dificuldades.

Aos meus pais, Antônio Mendonça e Aparecida Mendonça, pelo amor, incentivo e apoio incondicional.

À minha namorada, Gabriela Gênova, pela paciência, compreensão e apoio nos momentos mais delicados.

Ao meu orientador, Prof. Dr. Javam Machado, por me orientar e aconselhar em minha carreira acadêmica.

Ao Prof. Dr. José Maria Monteiro e Prof. Dr. Luciano Barbosa por, generosamente, aceitar meu convite e compor a banca.

À banca e amigos pelo trabalho e empenho que tiveram em revisar este trabalho.

Aos amigos Artur Barbosa, Bruno Leal, Daniel Praciano, Davi Torres, Eduardo Rodrigues, Edvar Filho, Felipe Timbó, Iago Chaves, Isabel Lima, Israel Vidal e Leonardo Linhares pelo companheirismo ao longo do desenvolvimento deste trabalho.

A todos os colegas do Laboratório de Sistemas e Banco de Dados (LSBD) pela agradável convivência diária.

A todos os familiares que, nos momentos de minha ausência, sempre continuaram a me incentivar e apoiar.

À Universidade Federal do Ceará (UFC) pelo ambiente criativo e amigável que proporciona.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo aporte financeiro para a viabilização deste trabalho.

Ao LSBD por ter proporcionado uma estrutura adequada para o desenvolvimento deste trabalho e aporte financeiro para eventos científicos.

A todos que direta ou indiretamente fizeram parte da minha formação.

“O insucesso é apenas uma oportunidade para recomeçar com mais inteligência.”

(Henry Ford)

RESUMO

A Privacidade Diferencial é um modelo matemático desenvolvido para dificultar o processo de identificação de indivíduos em conjuntos de dados estatísticos mantendo, ainda, a utilidade dos dados elevada. Embora a Privacidade Diferencial tenha sido amplamente utilizada para proteger a privacidade dos indivíduos, ela não foi desenvolvida para prover as mesmas garantias sobre dados correlacionados, uma vez que o modelo considera, em essência, a independência dos dados entre si. As técnicas existentes que utilizam Privacidade Diferencial em dados correlacionados buscam utilizar parâmetros de correlação ou coeficientes de correlação, e.g. *Pearson* e *Spearman*, para medir a correlação entre os indivíduos do conjunto de dados. No entanto, tais parâmetros e coeficientes tendem a introduzir, nas respostas das consultas, uma quantidade de ruído maior que a necessária, reduzindo consideravelmente a utilidade dos dados providos. Diferente dos trabalhos existentes, este trabalho propõe uma abordagem que agrupa os indivíduos semelhantes, i.e. aqueles com maior probabilidade de estarem correlacionados, através de duas técnicas de agrupamento: o Agrupamento Espacial Baseado em Densidade de Aplicações com Ruído (DBSCAN) e o Modelo de Mistura de Gaussianas (GMM). A abordagem também emprega o mecanismo de *Laplace*, que computa o ruído a ser adicionado nas respostas anonimizadas, satisfazendo, assim, as propriedades da Privacidade Diferencial. Os resultados da avaliação experimental confirmam os benefícios da estratégia de agrupamento em termos de eficácia, para melhoramento da utilidade, e desempenho comparado aos trabalhos existentes.

Palavras-chave: Privacidade de Dados. Privacidade Diferencial. Dados Correlacionados. Agrupamento de Dados. DBSCAN. GMM.

ABSTRACT

Differential Privacy is a mathematical model designed to hinder the process of distinguishing individuals' records on statistical databases, while maximizing data utility. Although Differential Privacy has been widely used for protecting the privacy of individual users' data, it was not designed to provide its guarantees for correlated data, since it considers, in essence, independence of records in the database. Existing techniques using Differential Privacy on correlated data attempt to use dependence parameters or correlation coefficients (such as Pearson or Spearman's Rank) to measure the correlation among records in a dataset. However, they tend to introduce an amount of noise higher than the necessary in the query answer, decreasing the data utility. Different from the existing works, we propose an approach that clusters similar records, which are more likely to be correlated, based on Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Gaussian Mixture Model (GMM). Our approach also employs a correlated Laplace mechanism to compute the privatized answers, satisfying the privacy guarantees of Differential Privacy. The experimental evaluation exhibits the benefits of our clustering strategy in terms of effectiveness and efficiency, considering data utility and privacy.

Keywords: Data Privacy. Differential Privacy. Correlated Data. Clustering. DBSCAN. GMM

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ALGORITMOS

LISTA DE ABREVIATURAS E SIGLAS

BIC	*Bayesian Information Criterion* / Critério de Informação Bayesiano
DBSCAN	*Density-Based Spatial Clustering of Applications with Noise* / Agrupamento Espacial Baseado em Densidade de Aplicações com Ruído
DC	*Density-Connected* / Conectado por Densidade
DDR	*Directly Density-Reachable* / Diretamente Alcançável por Densidade
DR	*Density-Reachable* / Alcançável por Densidade
EMD	*Earth Mover Distance*
EM	*Expectation Maximization* / Maximização de Expectativa
GMM	*Gaussian Mixture Model* / Modelo de Mistura de Gaussianas
MLE	*Maximum Likelihood Estimation* / Máxima Verossimilhança
RFE	*Recursive Feature Elimination* / Eliminação Recursiva de Atributo
RF	*Random Forest* / Floresta Aleatória
SVM	*Support Vector Machine* / Máquina de Vetor de Suporte
CPF	Cadastro de Pessoas Físicas
ER	Erro Relativo
ERM	Erro Relativo Médio
PD	Privacidade Diferencial

LISTA DE SÍMBOLOS

ε	Coeficiente de privacidade do modelo Privacidade Diferencial
M	Mecanismo do modelo Privacidade Diferencial
S	Saída de M
D	Conjunto de dados
D_i, D_j	Conjuntos de dados vizinhos
$D*$	Conjunto de dados com atributos selecionados
f	Função de consulta (*query*)
$f(D)$	Resultado de f aplicada sobre D
$f'(D)$	Resultado anonimizado de f aplicada sobre D
Δf	Sensibilidade de f
$\mathscr{D}$	Domínio de todos os conjuntos de dados
x	Variável aleatória (*Laplace*)
μ	Média da distribuição (*Laplace*)
b	Escala da distribuição (*Laplace*)
r, r_i, r_j	Indivíduos
δ_{ij}	Grau de correlação entre r_i e r_j
δ_0	Limiar do grau de correlação
Δ	Matriz de graus de correlação
D^i	Conjunto de dados com r_i
D^{-i}	Conjunto de dados sem r_i
CS_i	Sensibilidade de r_i
CS_f	Sensibilidade correlacionada de f
p, q	Pontos (*DBSCAN*)
C	*Cluster* (*DBSCAN*)
eps	Raio (*DBSCAN*)
$minPoints$	Número mínimo de pontos em C (*DBSCAN*)

$\mathscr{M}$	Modelo (*GMM*)
ω_{ij}	Probabilidade de r_i e r_j pertencerem ao mesmo *cluster* em $\mathscr{M}$
Ω	Matriz de probabilidades

SUMÁRIO

1 INTRODUÇÃO

1.1 Motivação

O grande volume de dados produzidos e coletados diariamente por meio de aplicações, sejam através de dispositivos móveis, sensores, aplicações web, entre outros, torna possível a realização de análises sobre os indivíduos envolvidos. A análise sobre os dados pode ser designada para os mais diversos fins, tais como: entendimento do comportamento humano, descoberta de padrões, identificação de tendências, análises estatísticas e muito mais. Entretanto, o processo de análise sobre os dados pode acabar violando o direito de privacidade dos indivíduos, uma vez que pode haver a necessidade de acessar informações sensíveis dos mesmos.

Chamamos de *dataholders* aqueles que detêm o controle sobre os dados, como organizações e provedores de serviço. Diante de processos de coleta e análise de dados via web, os *dataholders* passam a enfrentar desafios adicionais. Uma vez que o acesso indiscriminado aos dados pode levar à violação da privacidade dos indivíduos, é preciso que o conjunto de dados atenda previamente a alguns critérios de privacidade antes de sua liberação para os devidos fins de análise. Uma possível solução para o problema seria modificar o conjunto de dados original, através de um algoritmo de anonimização, antes de publicá-lo. Por exemplo, um conjunto de dados contendo informações sobre os salários de indivíduos poderia ter seus valores exatos generalizados em intervalos, de maneira que todos os valores de salários acima de R$1.000,00 e abaixo de R$5.000,00 fossem representados por [R$1.000,00 - R$5.000,00]. Outra possível solução seria utilizar o mesmo algoritmo de anonimização, mas agora para publicar apenas informações estatísticas, por meio de consultas, sobre os dados. Por exemplo, utilizando o mesmo conjunto de dados de salários, um usuário que submete uma consulta interessado em descobrir o número de indivíduos que possuem salário entre R$1.000,00 e R$5.000,00 obteria, como resposta, o valor exato da consulta acrescido de algum ruído. O primeiro cenário ilustra os modelos de privacidade sintáticos, enquanto que o segundo os interativos.

A aplicação de um modelo de privacidade sobre um conjunto de dados é imprescindível para evitar que indivíduos sejam re-identificados quando os dados se tornam públicos. Todavia, todo modelo de privacidade acaba provocando mudanças nas informações publicadas, afetando diretamente a qualidade do processo de análise sobre os dados, ou seja, diminuindo sua utilidade. Portanto, gerenciar esse *trade-off* entre privacidade e utilidade se torna um outro grande desafio.

Embora todos tenham o mesmo objetivo de atender o propósito de garantir a privacidade dos indivíduos, modelos sintáticos e interativos possuem algumas especificidades. Modelos de privacidade sintáticos consideram o conhecimento prévio que algum adversário (usuário malicioso) possa ter sobre os dados. É necessário que os critérios de privacidade de cada algoritmo de anonimização sejam definidos cautelosamente, visto que ter conhecimento prévio sobre os dados aumenta substancialmente a chance de descoberta de algum indivíduo. Em contrapartida, modelos de privacidade interativos buscam eliminar essa limitação, propondo soluções que independem do conhecimento prévio de um adversário. Um dos modelos de privacidade interativo mais conhecidos é a Privacidade Diferencial (PD).

O modelo de Privacidade Diferencial (DWORK, 2006) surgiu nas últimas décadas como um modelo matemático rigoroso que provê fortes garantias de privacidade. O modelo recebeu considerável atenção da comunidade de privacidade devido às suas características inerentes de ser independente de qualquer conhecimento adversário e, até mesmo, de poder computacional (DWORK *et al.*, 2014; LEE; CLIFTON, 2011). A Privacidade Diferencial garante que qualquer sequência de respostas de consultas é igualmente possível de ocorrer independente da presença, ou ausência, de qualquer indivíduo no conjunto de dados. Em outras palavras, a adição, ou remoção, de um indivíduo do conjunto de dados não irá afetar substancialmente o resultado de qualquer análise estatística realizada sobre o conjunto de dados. Por exemplo, modificar o conjunto de dados em apenas um indivíduo não ocasionará grandes mudanças nos resultados das consultas de maneira a comprometer as análises estatísticas sobre os dados.

Embora o modelo de Privacidade Diferencial venha sendo amplamente utilizado como um modelo robusto para proteger a privacidade dos dados dos indivíduos, o modelo não foi desenvolvido, em essência, para prover as mesma garantias de privacidade sobre dados correlacionados. A Privacidade Diferencial assume, implicitamente, que todos os registros em um conjunto de dados são independentes. De fato, essa suposição é correta em muitos casos. Entretanto, em aplicações do mundo real podem existir fortes evidências de relacionamento entre os registros. Por exemplo, um conjunto de dados de rede social contendo relacionamentos de amizade entre os usuários, representado por nós e arestas, mantém informação relevante de relacionamento entre usuários, representada explicitamente por meio das arestas. Uma vez que existem essas relações entre diferentes usuários, e considerando que dois itens relacionados são propícios a possuírem vários outros itens também relacionados em comum, um usuário malicioso é capaz de deduzir a existência de correlação, i.e., ligações entre os usuários, mesmo

que, inicialmente, estes não estivessem conectados, podendo levar à violações de privacidade dos usuários. Em outro exemplo, podemos assumir que informações privadas podem ser inferidas ao se utilizar informações de conhecimento público compartilhadas por usuários que compartilham informações semelhantes. Além disso, um usuário malicioso é capaz de deduzir a suscetibilidade de um usuário à uma determinada doença altamente contagiosa ao ter conhecimento que os familiares desse mesmo usuário estão contidos no mesmo conjunto de dados (CHEN *et al.*, 2014; KIFER; MACHANAVAJJHALA, 2011). Portanto, aplicar o modelo de Privacidade Diferencial para proteger a privacidade dos indivíduos no contexto de dados correlacionados é uma questão que precisa ser abordada.

Nos últimos anos, alguns autores têm estudado o modelo de Privacidade Diferencial para cenários onde os dados estão correlacionados. Kifer et al. (KIFER; MACHANAVAJJHALA, 2014) definiu um *framework* de privacidade, denominado *Pufferfish*, o qual permite que especialistas no domínio de aplicação customizem novas definições de privacidade. Embora esse *framework* possa ser empregado sobre dados correlacionados, a abordagem não é tão trivial de ser implementada, uma vez que é necessário definir um conjunto de requisitos para proteger os registros correlacionados contra usuários maliciosos, denominados segredos potenciais, pares discriminativos e cenários de evolução. Liu et al. (LIU *et al.*, 2016) formalizou a noção de Privacidade Diferencial Dependente, a qual incorpora dependências probabilísticas entre os relacionamentos existentes no conjunto de dados através de parâmetros de dependência. Entretanto, a abordagem requer que os parâmetros de dependência sejam conhecidos ou, pelo menos, muito bem estimados. Além disso, Zhu et al. (ZHU *et al.*, 2015) propôs uma maneira de aplicar o modelo de Privacidade Diferencial sobre conjuntos de dados correlacionados, que considera a correlação entre os registros do conjunto de dados. Tal correlação é mensurada através do Coeficiente de Correlação de *Pearson*. Embora os trabalhos de Zhu et al. (ZHU *et al.*, 2015) satisfaçam as propriedades da Privacidade Diferencial, o método nele proposto para mensurar a correlação resulta em uma inserção de ruído maior que a necessária, reduzindo a utilidade dos dados.

Neste trabalho, apresentamos uma estratégia de Privacidade Diferencial que emprega técnicas de agrupamento para identificar os relacionamentos existentes entre itens de um conjunto de dados. A correlação entre os indivíduos que compõem os relacionamentos é medida de acordo com a disposição dos grupos identificados. As técnicas de agrupamento foram empregadas na solução por serem algoritmos amplamente conhecidos e utilizados com o objetivo de agrupar

conjuntos de objetos similares entre si. Portanto, consideramos que a utilização de técnicas de agrupamento são mais eficientes para identificar os relacionamentos entre indivíduos no processo de garantia de privacidade por meio da Privacidade Diferencial.

1.2 Objetivos

1.2.1 Objetivo geral

Diante do cenário apresentado na motivação, o objetivo geral deste trabalho consiste em produzir uma solução interativa, através do modelo de Privacidade Diferencial, que preserve a privacidade de indivíduos em conjuntos de dados onde evidencia-se a existência de correlação entre os indivíduos, enquanto se mantém a utilidade dos dados e o bom desempenho da solução.

1.2.2 Objetivos específicos

Com o objetivo de atender ao objetivo geral deste trabalho, estabelecemos os seguintes objetivos específicos:

- Definir uma estratégia para identificar relacionamentos implícitos em um conjunto de dados;
- Mensurar a correlação entre os indivíduos de maneira acurada a fim de minimizar a perda de informação gerada enquanto se mantêm as propriedades do modelo de privacidade;
- Avaliar a eficiência da solução proposta utilizando dados tabulares reais em termos de utilidade e tempo de execução.

1.3 Contribuições

Como resultado desta dissertação, implementamos um algoritmo que tem como objetivo prover respostas de consultas anonimizadas aplicando o modelo de Privacidade Diferencial sobre dados correlacionados, visando preservar o máximo de utilidade possível. Nós assumimos que a perda de informação é medida em função do erro relativo, o qual avalia, em termos percentuais, o quão distantes das respostas originais as respostas anonimizadas estão.

Em particular, as principais contribuições deste trabalho são:

- Uma estratégia de agrupamento para identificar relacionamentos implícitos em

dados correlacionados utilizando os algoritmos *Density-Based Spatial Clustering of Applications with Noise* / Agrupamento Espacial Baseado em Densidade de Aplicações com Ruído (*DBSCAN*) e *Gaussian Mixture Model* / Modelo de Mistura de Gaussianas (*GMM*);

- Uma estratégia acurada para mensurar os graus de correlação entre os indivíduos;
- Uma estratégia de anonimização para publicação de consultas que adota o modelo de Privacidade Diferencial com o objetivo de minimizar a perda de informação e melhorar o desempenho.

1.3.1 Produção científica

Os resultados parciais da pesquisa realizada nesta dissertação foram publicados no seguinte artigo:

- André L. C. Mendonça, Felipe T. Brito, Leonardo S. Linhares, and Javam C. Machado. DiPCoDing: A Differentially Private Approach for Correlated Data with Clustering. In Proceedings of the 21st International Database Engineering & Applications Symposium. (IDEAS 2017)

Nova submissão para um periódico internacional com a descrição completa da pesquisa e seus resultados está em fase de elaboração. Também no contexto deste trabalho, mas de maneira indireta, o seguinte artigo foi publicado:

- Felipe T. Brito, Antônio C. Araújo Neto, Camila F. Costa, André L.C. Mendonça, Javam C. Machado, A Distributed Approach for Privacy Preservation in the Publication of Trajectory Data. The Workshop on Privacy in Geographic Information Collection and Analysis. (GeoPrivacy@ACM SIGSPATIAL 2015)

1.4 Estrutura da dissertação

Esta dissertação está organizada da seguinte forma: No Capítulo 2 são apresentados conceitos e definições fundamentais sobre preservação de privacidade destacando, principalmente, o modelo de Privacidade Diferencial. Também são apresentadas algumas técnicas de agrupamento, evidenciando sua ligação com a existência de correlação nos dados. O Capítulo 3 ressalta e discute os trabalhos relacionados mais relevantes no contexto de Privacidade Diferencial sobre dados correlacionados e, também, sobre técnicas de agrupamento. Em seguida, o

Capítulo 4 apresenta a nossa solução proposta, dividida em três abordagens, a qual utiliza as técnicas de agrupamento *DBSCAN* e *GMM* para identificar os indivíduos correlacionados em um conjunto de dados e mensura seus respectivos graus de correlação através de uma estratégia acurada e eficiente. Ainda no capítulo, apresentamos os resultados obtidos por meio de uma série de experimentos sobre conjuntos de dados reais. Finalmente, o Capítulo 5 conclui o trabalho apresentando um resumo dos resultados alcançados e mostrando direções de pesquisas futuras.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo consiste na fundamentação teórica necessária para o entendimento deste trabalho, incluindo os problemas conhecidos na literatura e as técnicas utilizadas para o desenvolvimento da solução proposta. Uma visão geral acerca da privacidade de dados é apresentada na Seção 2.1, destacando sua necessidade e importância. Na Seção 2.2 apresentamos os modelos de privacidade mais conhecidos na literatura, com ênfase no modelo Privacidade Diferencial, que tem servido de base para a realização da pesquisa. Por fim, apresentamos em detalhes, na Seção 2.3, as técnicas de agrupamento empregadas na solução proposta neste trabalho.

2.1 Privacidade de Dados

Grandes volumes de dados têm sido coletados por governos, corporações e instituições ao redor do mundo. Dados são extremamente valiosos para os diversos tipos de organizações envolvidas nesse processo devido às suas inúmeras finalidades. Por exemplo, seria de grande valia que empresas de publicidade tivessem conhecimento sobre padrões de navegação de usuários para que pudessem oferecer propagandas de acordo com cada perfil de usuário. Instituições bancárias poderiam entender como seus clientes utilizam seus cartões de crédito e, dessa forma, oferecer outros tipos de serviços. Diversas outras redes, tais como: supermercados, farmácias e restaurantes, também poderiam entender o comportamento de seus clientes para que pudessem oferecer os produtos e serviços mais requisitados e, assim, diminuir possíveis prejuízos, como grandes quantidades de produtos parados em estoque, e aumentar a rotatividade de clientes. Dados de mobilidade urbana poderiam ser empregados para identificar e prever vias congestionadas, melhorar o transporte público e planejar obras de infraestrutura, assim como dados de criminalidade poderiam ser empregados para prever áreas mais críticas, que carecem de segurança, e assim facilitar a realização de remanejamento eficaz das forças de segurança. Esses são apenas alguns exemplos que ilustram a importância dos dados. Em suma, a análise sobre os dados pode ser designada para os mais diversos fins, tais como: entendimento do comportamento humano, descoberta de padrões, identificação de tendências, análises estatísticas e muito mais.

No entanto, a publicação dos dados para análises e descoberta de padrões pode acabar colocando em risco a privacidade dos indivíduos envolvidos. Se os dados coletados forem mantidos em seu formato original e, por acaso, esses dados acabem caindo em mãos de

usuários maliciosos (também conhecidos por adversários ou atacantes), os indivíduos podem ter suas identidades facilmente identificadas, violando, assim, sua privacidade. Portanto, na tentativa de impedir que os indivíduos sejam identificados após a publicação dos dados, todos os identificadores explícitos devem ser removidos ou suprimidos previamente. Identificadores explícitos são aqueles que identificam um indivíduo de maneira unívoca, como, por exemplo, nome, e-mail e Cadastro de Pessoas Físicas (CPF). No entanto, mesmo com a remoção prévia dos identificadores explícitos, usuários maliciosos ainda são capazes de descobrir informações sensíveis sobre os indivíduos através de seus semi-identificadores. Semi-identificadores são dados que podem ser combinados com informações externas, ou de conhecimento prévio de um usuário malicioso, para identificar os indivíduos. Assim, um usuário malicioso é capaz de descobrir o registro, em um conjunto de dados, que contém informações acerca de um determinado indivíduo com alta probabilidade. Dessa forma, caso os dados sejam publicados sem as devidas precauções de privacidade, poderão ocorrer diversas violações de privacidade, uma vez que esses dados podem fornecer informações sensíveis sobre os indivíduos, tais como: salário, doenças, crenças, preferências sexuais, entre outras.

A Tabela 1 apresenta um exemplo de conjunto de dados publicado por uma empresa de Tecnologia da Informação. Identificadores explícitos, como “Nome” e “CPF”, foram removidos previamente da tabela. A coluna “Data de Nascimento” representa os semi-identificadores, uma vez que podem ser combinados com informações externas para tornar possível a identificação de algum indivíduo. Já as colunas “Cargo” e “Salário” representam os atributos sensíveis.

Tabela 1 – Exemplo de um conjunto de dados publicado por uma empresa

ID	**Data de Nascimento**	**Cargo**	**Salário (R$)**
1	01/01/1980	Gerente	15.000,00
2	02/02/1985	Analista	9.000,00
3	03/03/1990	Analista	7.500,00
4	04/04/1995	Trainee	4.000,00
5	05/05/2000	Estagiário	1.000,00

Garantir a privacidade dos indivíduos tem impacto direto na qualidade dos dados publicados, tendo em vista que a utilidade e a privacidade dos dados são duas grandezas inversamente proporcionais. Dessa forma, quanto mais privados os dados são, mais distantes de suas representações originais estão, reduzindo a expressividade das informações providas, ou seja, reduzindo sua utilidade. De maneira análoga, quanto menos privados os dados são, mais próximos de suas representações originais estão, aumentando a expressividade das informações

providas e, também, a utilidade dos dados. Entretanto, quanto menos privados, maiores são as chances de usuários maliciosos violarem a privacidade dos indivíduos. A Figura 1 ilustra bem esse comportamento. Como mencionado, temos que quanto maior a utilidade dos dados, menor a privacidade, e vice-versa. Portanto, manter a utilidade dos dados e, ao mesmo tempo, garantir a privacidade dos indivíduos se torna um problema bastante complexo.

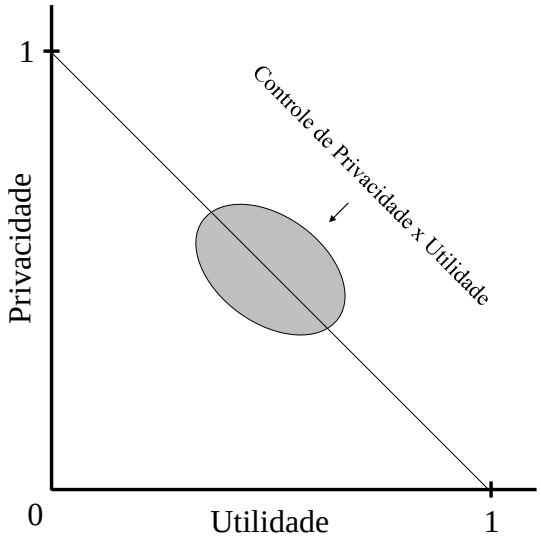


Figura 1 – Balanceamento entre utilidade e privacidade.

2.2 Modelos de Privacidade

Diversos modelos de privacidade de dados têm sido propostos para tentar resolver esse problema. Atualmente, os modelos são classificados como sintáticos ou interativos. Em modelos sintáticos, os dados devem obedecer a algum critério antes de serem publicados. Por sua vez, os modelos interativos são responsáveis por disponibilizar resultados de consultas, com base em algum modelo matemático responsável por prover informações estatísticas acerca do conjunto de dados original.

Como visto na seção anterior, os dados precisam ser modificados antes de serem disponibilizados para que os indivíduos envolvidos não tenham sua privacidade violada, e assim tenham suas informações sensíveis descobertas por usuários maliciosos. Pesquisas têm mostrado que anonimizar os dados antes de disponibilizá-los publicamente, ou para terceiros, se trata da estratégia mais promissora para garantir a privacidade dos indivíduos (FUNG *et al.*, 2010; WILLISON *et al.*, 2008). A anonimização consiste em um grupo de técnicas que modifica um conjunto de dados original, transformando-o em um novo conjunto de dados anonimizado, de

forma que os dois conjuntos de dados não se assemelham entre si, mas possuem semântica e sintaxe semelhantes. As modificações realizadas pelo algoritmo de anonimização sobre o conjunto de dados têm o objetivo de garantir a privacidade dos indivíduos. Técnicas como: supressão, generalização e perturbação são algumas das mais utilizadas nessa tarefa de modificar os dados. Dessa forma, as propriedades específicas de cada modelo de privacidade de dados devem ser garantidas por meio de algum algoritmo de anonimização.

As Tabela 2 apresenta um exemplo de conjunto de dados hipotético e a Tabela 3 apresenta uma possível versão anonimizada do mesmo conjunto, após ser modificada por técnicas de supressão e generalização. Como observado na versão anonimizada, os atributos "Nome" e "CPF" foram suprimidos por se tratarem de identificadores explícitos e tiveram seu valores substituídos pelo símbolo de "*". Por sua vez, os atributos "Idade" e "Profissão" tiveram seus valores generalizados. O atributo "Idade" foi generalizado em diferentes intervalos, [25-35] ou [40-55], enquanto o atributo "Profissão" foi generalizado baseando-se em uma possível hierarquia, identificado-se duas possíveis áreas de profissão, Saúde ou TI.

Tabela 2 – Conjunto de dados hipotético.

Nome	**CPF**	**Idade**	**Profissão**
Ana	111.111.111-99	25	Enfermeira
Bárbara	222.222.222-99	33	Médica
Celso	333.333.333-99	28	Programador
Diogo	444.444.444.99	42	Dentista
Emerson	555.555.555.99	51	Analista de TI

Tabela 3 – Conjunto de dados anonimizado após operações de supressão e generalização.

Nome	**CPF**	**Idade**	**Profissão**
*	*	[25-35]	Saúde
*	*	[25-35]	Saúde
*	*	[25-35]	TI
*	*	[40-55]	Saúde
*	*	[40-55]	TI

Diante dos problemas decorridos pela publicação indevida de dados, no que diz respeito à privacidade dos indivíduos, os *dataholders* precisam tomar algumas decisões antes de disponibilizarem os dados para fins de análise. Uma delas consiste em decidir o modelo de privacidade que será aplicado sobre os dados, de forma que os dados disponibilizados ainda sejam úteis. Embora os modelos de privacidade, sejam eles sintáticos ou interativos, possuam objetivos em comum, como o de garantir a privacidade dos indivíduos, também possuem algumas

características específicas que os diferem uns dos outros.

Os modelos de privacidade costumam garantir a privacidade dos indivíduos ao impedir que usuários maliciosos sejam capazes de realizar ataques com a intenção de violar a privacidade desses indivíduos. Em modelos sintáticos, os ataques geralmente são divididos em: ataque de ligação ao registro, ataque de ligação ao atributo e ataque de ligação à tabela. Já em modelos interativos, busca-se impedir o ataque probabilístico. Cada tipo de ataque mencionado é definido por:

- **Ataque de Ligação ao Registro**: o usuário malicioso é capaz de identificar o registro de um determinado indivíduo, cujas informações estão presentes no conjunto de dados publicado;
- **Ataque de Ligação ao Atributo**: nesse tipo de ataque o usuário malicioso é capaz de inferir os valores de atributos sensíveis pertencentes a um determinado indivíduo, baseando-se no conjunto de atributos sensíveis associados ao grupo no qual o indivíduo pertence. No entanto, o usuário malicioso não é capaz de identificar o registro completo desse indivíduo;
- **Ataque de Ligação à Tabela**: diferentemente dos tipos de ataques anteriores, nos quais o usuário malicioso tem pleno conhecimento de que as informações referentes a um determinado indivíduo foram publicadas, nesse tipo de ataque o usuário malicioso está interessado em inferir com convicção se um determinado indivíduo está presente, ou não, no conjunto de dados publicado;
- **Ataque Probabilístico**: por fim, esse tipo de ataque enfatiza como um usuário malicioso mudaria seu pensamento probabilístico acerca de um determinado indivíduo, uma vez que esse usuário teve acesso ao conjunto de dados publicado. O usuário malicioso não está interessado em utilizar o conhecimento que ele detém sobre o indivíduo e associá-lo para identificar o registro, os valores de atributos sensíveis ou a participação desse indivíduo no conjunto de dados.

Vistos os tipos de ataques que podem ser praticados por usuários maliciosos, entraremos em detalhes nos modelos de privacidade de dados. Primeiramente, abordaremos, de forma resumida, os modelos sintáticos, destacando alguns dos mais conhecidos da literatura, como: k-anonimato, l-diversidade, t-proximidade e δ-presença, e sempre mencionando os tipos de ataque que cada modelo se propõe a prevenir. Por fim, abordaremos os modelos interativos, descrevendo com detalhe o modelo de Privacidade Diferencial, o qual faz parte da solução proposta neste trabalho.

2.2.1 Modelos Sintáticos

2.2.1.1 k-anonimato

Considerado o mais conhecido dos modelos de privacidade, o *k*-anonimato foi proposto por (SWEENEY, 2002) com o objetivo de prevenir ataques de ligação ao registro. Esse modelo assegura que, para cada combinação de *k* semi-identificadores, existem, pelo menos, *k* registros distintos no conjunto de dados publicado, formando uma classe de equivalência. O *k*-anonimato atua, portanto, sobre o princípio da indistinguibilidade, onde é garantido que cada registro em um conjunto de dados *k*-anônimo, i.e., que satisfaz as propriedades do modelo *k*-anonimato, é indistinguível a, no mínimo, outros $k-1$ registros no conjunto de dados, em relação ao conjunto de semi-identificadores. Portanto, a probabilidade de um usuário malicioso violar a privacidade de um indivíduo através de um ataque de ligação ao registro não poderá ser maior do que $\frac{1}{k}$. O parâmetro *k* do modelo é responsável por balancear a utilidade e a privacidade dos dados. Assim, quanto maior o valor de *k*, maior será a privacidade dos dados e, consequentemente, menor a utilidade dos dados, e vice-versa. É importante ressaltar que não existem abordagens analíticas para determinar um valor ótimo para o parâmetro *k* (DEWRI *et al.*, 2008). Além disso, encontrar um valor ótimo para o parâmetro *k* é NP-difícil (MEYERSON; WILLIAMS, 2004). Dessa forma, cabe aos *dataholders* a difícil tarefa de determinar o valor do parâmetro *k* quando da aplicação do processo de anonimização por k-anonimato sobre um conjunto de dados.

2.2.1.2 l-diversidade

Uma vez que o modelo *k*-anonimato, apresentado anteriormente, não é capaz prevenir ataques de ligação ao atributo, o modelo *l*-diversidade foi proposto por (MACHANAVAJJHALA *et al.*, 2006) com o objetivo de suprir essa limitação. O modelo assegura que para cada classe de equivalência, existem, pelo menos, *l* valores distintos para cada atributo sensível. Dessa forma, um usuário malicioso detentor de conhecimento prévio poderá apenas descobrir a classe de equivalência de um determinado indivíduo, sendo incapaz de inferir os valores de atributos sensíveis pertencentes ao indivíduo com probabilidade maior que $\frac{1}{l}$.

2.2.1.3 t-proximidade

O modelo *t*-proximidade, proposto por (LI *et al.*, 2007), corrige algumas limitações do modelo *l*-diversidade, em específico à vulnerabilidade ao *Skewness Attack*, além de também previnir ataques de ligação ao atributo. Em um *Skewness Attack*, um usuário malicioso é capaz de inferir as informações sobre os atributos sensíveis de um indivíduo a partir do conhecimento da frequência de ocorrência desses atributos sensíveis. Dessa forma, o modelo *t*-proximidade assegura que a frequência de ocorrência de cada atributo sensível em cada classe de equivalência seja semelhante à sua distribuição global. O parâmetro t define a distância máxima permitida entre a distribuição das classes de equivalência e a distribuição global. Seu valor pode ser mensurado através da *Earth Mover Distance* (*EMD*) (LI *et al.*, 2007), uma métrica bastante utilizada para medir a distância entre duas distribuições, assumindo valores no intervalo $[0,1]$.

2.2.1.4 δ-presença

O δ-presença está contido nos modelos de privacidade que garantem a privacidade dos indivíduos ao prevenir ataques de ligação à tabela. Proposto por (NERGIZ *et al.*, 2007), o modelo assegura que um usuário malicioso é incapaz de deduzir se um indivíduo está presente, ou não, no conjunto de dados. O parâmetro δ representa o limite, através da forma $\delta = (\delta_{min}, \delta_{max})$, de confiança que um usuário malicioso possui para inferir a presença, ou ausência, de um indivíduo no conjunto de dados publicado.

2.2.2 Modelos Interativos

2.2.2.1 Privacidade Diferencial

Vimos que nos modelos de privacidade sintáticos apresentados anteriormente um conjunto de dados é previamente modificado, através de um algoritmo de anonimização, para publicação com garantias de privacidade dos indivíduos representados no conjunto. Quando existente, a violação de privacidade dos indivíduos ocorre quando um usuário malicioso, juntamente com seu conhecimento prévio adquirido, se torna capaz de identificar os indivíduos. Sob outra visão, o modelo de Privacidade Diferencial objetiva publicar resultados de consultas, e não o conjunto de dados, de forma que ruídos são adicionados aos resultados das consultas. Em suma, os dados são perturbados para garantir a privacidade dos indivíduos, fazendo com que um

usuário malicioso não consiga concluir algo com 100% de certeza. Além disso, as informações providas pelo modelo não mais remetem a um indivíduo em particular, mas sim ao conjunto de dados em sua totalidade. Dessa forma, o modelo em questão foi proposto para impedir ataques probabilísticos.

2.2.2.1.1 Conceitos Básicos

Proposta por (DWORK, 2006), a Privacidade Diferencial (PD) consiste em um modelo matemático que oferece sólidas garantias de privacidade. O modelo tem como objetivo fornecer informações estatísticas sobre um conjunto de dados sem comprometer a privacidade dos indivíduos envolvidos, através de um algoritmo aleatório, geralmente chamado de mecanismo. Esse mecanismo é responsável por introduzir aleatoriedade e proteger os resultados das consultas sobre o conjunto de dados. Inicialmente, a Privacidade Diferencial foi projetada para atuar em um ambiente interativo, onde usuários submetem consultas a um conjunto de dados e o ambiente, por sua vez, responde às consultas através de um algoritmo de anonimização. No entanto, mostrou-se que também é possível aplicar o modelo sobre um conjunto de dados de forma a publicar sua versão anonimizada, ou seja, com os dados perturbados. Entretanto, o foco deste trabalho está voltado para a aplicação da Privacidade Diferencial em ambientes interativos. A Figura 2 apresenta o fluxo em um ambiente interativo no modelo de Privacidade Diferencial. Nela um usuário submete uma consulta consulta a um conjunto de dados, a resposta da consulta da consulta é computada e, por fim, um mecanismo introduz uma quantidade de ruído à resposta real da consulta e a retorna ao usuário.

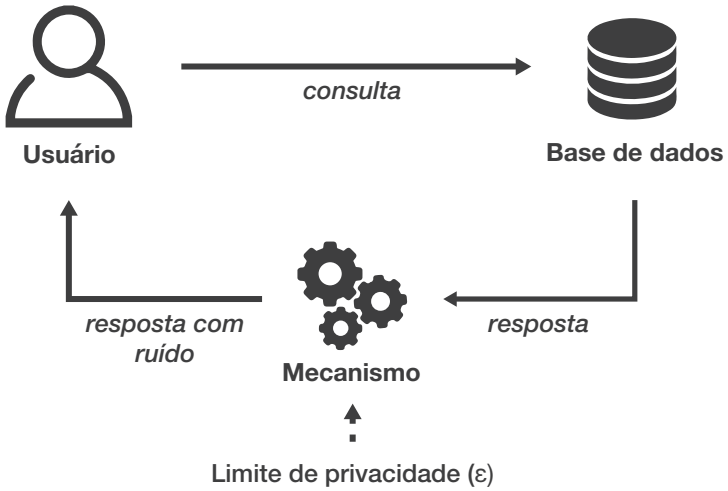


Figura 2 – Ambiente interativo no modelo de Privacidade Diferencial.

A Privacidade Diferencial assegura que qualquer sequência de resultados, i.e., respostas de consultas, é igualmente possível de ocorrer e independe da presença, ou ausência, de qualquer indivíduo no conjunto de dados [19]. Dessa forma, a adição ou remoção de algum indivíduo não irá afetar drasticamente o resultado de qualquer análise estatística realizada sobre o conjunto de dados. Portanto, o modelo parte do pressuposto de que um usuário malicioso não deve ser capaz de aprender nada sobre um determinado indivíduo, de forma que ele já não poderia ter aprendido antes sem ter acesso ao conjunto de dados.

O exemplo a seguir motiva o a necessidade do modelo de Privacidade Diferencial ao apresentar uma violação de privacidade. Suponha que a altura dos indivíduos represente informações muito sensíveis, visto que disponibilizá-las poderia causar uma série de violações na privacidade dos indivíduos. Acontece que uma empresa atuante na coleta de informações demográficas acaba de disponibilizar um conjunto de dados, para consultas, que fornece informações acerca da altura média das mulheres de diferentes países. Um usuário malicioso que detém acesso aos resultados providos pelo conjunto de dados e à informação complementar de que Fred, o indivíduo o qual ele deseja descobrir informações, é 5 centímetros mais alto que a altura média da população feminina que reside na Lituânia, pode facilmente descobrir a altura real de Fred após requisitar uma consulta ao conjunto de dados que provê a altura média dessas mulheres. Portanto, o usuário malicioso foi capaz de violar a privacidade do indivíduo alvo ao determinar a sua altura com exatidão, mesmo que ele não estivesse presente no conjunto de dados, o que não teria sido possível caso o adversário não tivesse acesso ao conjunto de dados.

2.2.2.1.2 Definição Formal (DWORK, 2006)

Definição 1 *Dado um algoritmo aleatório (mecanismo) M, esse mecanismo garante o ε-Privacidade Diferencial se para todos os conjuntos de dados vizinhos D_1 e D_2, i.e., que diferem em no máximo um elemento (registro), e para todo S contido no conjunto de todos os resultados possíveis provenientes de M, isto é, para todo $S \subseteq Imagem(M)$, temos que:*

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

onde Prob é a probabilidade dada sobre a aleatoriedade de M.

Em suma, a definição do modelo afirma que, ao aplicar um mecanismo sobre dois conjuntos de dados vizinhos, a diferença entre as probabilidades do mecanismo retornar a mesma

resposta para uma mesma consulta é limitada pelo parâmetro ε. Assim, para quaisquer pares de conjuntos de dados vizinhos D_1 e D_2 e resultados provenientes do mecanismo, um usuário malicioso não será capaz de distinguir entre D_1 e D_2 baseando-se apenas na resposta fornecida pelo mecanismo.

Para facilitar o entendimento, a Figura 3 apresenta um exemplo de saída de um mecanismo M sobre os conjuntos de dados vizinhos D_1 e D_2 a partir de um determinado valor do parâmetro ε. As saídas de M são representadas por $f(D)+$ *ruído*, onde $f(D)$ representa o resultado de uma consulta f aplicada sobre um conjunto de dados D. No exemplo em questão, tem-se que os resultados seguem uma distribuição de *Laplace*. Nesse caso, o algoritmo de anonimização é denominado mecanismo de *Laplace*.

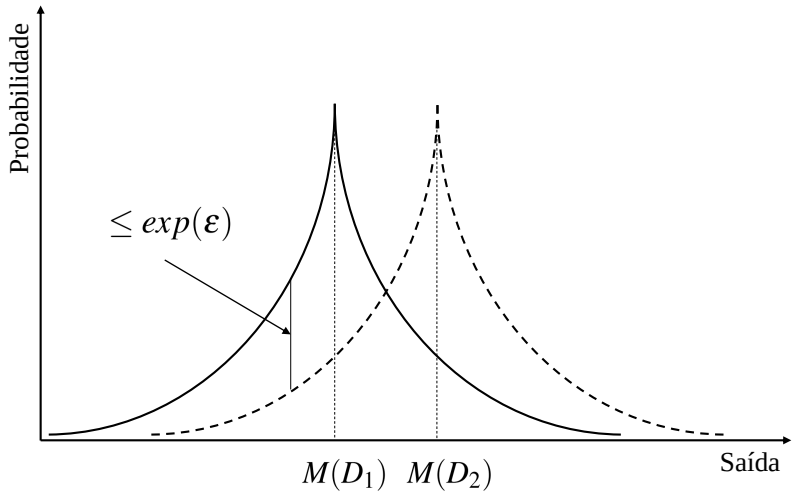


Figura 3 – Probabilidades de saída de um mecanismo M sobre os conjuntos de dados vizinhos D_1 e D_2.

É importante ressaltar que as garantias oferecidas pelo modelo de Privacidade Diferencial independem de qualquer poder computacional ou informações que um usuário malicioso possa ter obtido. Se um determinado indivíduo decide participar de um conjunto de dados, o qual será submetido a análises estatísticas através de um mecanismo ε-Diferencialmente Privado, tal mecanismo garante que não haverá um aumento significativo na probabilidade de violação de privacidade dos indivíduos se comparado com a probabilidade de quando o mesmo indivíduo decide por não participar do conjunto de dados. Portanto, o modelo oferece altas garantias de privacidade sobre a identidade dos indivíduos.

Diferentemente dos modelos apresentados anteriormente, como o k-anonimato, descrito na Seção 2.2.1.1, onde o parâmetro k é responsável por determinar o grau de privacidade desejado na publicação dos dados, o parâmetro ε do modelo de Privacidade Diferencial não

possui uma correlação explícita com a privacidade dos indivíduos. O valor de ε irá depender de alguns fatores, como: o tipo de consulta realizada, os próprios dados que compõem o conjunto de dados ou interesses dos próprios *dataholders*. Não existem abordagens especializadas em descobrir um valor ótimo para o parâmetro ε, sendo necessário, em alguns casos, encontrá-lo empiricamente. No entanto, é recomendado que o ε assuma valores pequenos, da ordem de 0.01, 0.1, 1.0, ln_2 ou ln_3 (DWORK, 2008). Quanto menor o valor de ε, maior a privacidade e menor a utilidade, e vice-versa.

2.2.2.1.3 Mecanismo Diferencial

Como mencionado nas seções anteriores, a Privacidade Diferencial foi definida, inicialmente, sob a perspectiva de um modelo interativo, onde os usuários submetem consultas a um conjunto e dados e recebem, por meio de um mecanismo, uma resposta ε-Diferencialmente Privada. Técnicas que utilizam esse modelo de privacidade tem o objetivo de criar um mecanismo que irá adicionar ruído à resposta de uma consulta realizada por um usuário, de maneira que esse ruído gerado seja independente do conjunto de dados.

A quantidade de ruído necessária irá depender do tipo de consulta realizada sobre o conjunto de dados. Assim, é necessário definir o que é a sensibilidade de uma consulta. Entretanto, é preciso entender, a priori, a noção de conjuntos de dados vizinhos que, embora já mencionada, será definida formalmente a seguir.

Definição 2 *Dado um conjunto de dados D, todos os conjuntos de dados D_i provenientes da remoção de algum indivíduo i pertencente ao conjunto de dados D são definidos como vizinhos.*

Considere a Tabela 4a, que representa um conjunto de dados *D* contendo os atributos "Estado", onde um indivíduo reside, e seu "Time do Coração", representando o time de futebol que o indivíduo torce. A Tabela 4b apresenta um dos possíveis conjuntos de dados vizinhos D_i gerado a partir do conjunto de dados *D* após a remoção do registro com o "ID" = 4.

Uma vez que a noção de conjuntos de dados vizinhos foi apresentada, podemos, agora, definir o que é a sensibilidade de uma consulta.

Definição 3 *Seja $\mathcal{D}$ o domínio de todos os conjuntos de dados e f uma função de consulta que mapeia conjuntos de dados a vetores de números reais. A sensibilidade da função f é dada por:*

Tabela 4 – Exemplo de conjuntos de dados vizinhos.

ID	Estado	Time do Coração
1	São Paulo	Palmeiras
2	São Paulo	Santos
3	Rio de Janeiro	Flamengo
4	Ceará	Flamengo
5	Ceará	Fortaleza

(a)

ID	Estado	Time do Coração
1	São Paulo	Palmeiras
2	São Paulo	Santos
3	Rio de Janeiro	Flamengo
5	Ceará	Fortaleza

(b)

$$\Delta f = \max_{D_i, D_j \in \mathscr{D}} \| f(D_i) - f(D_j) \|_1$$

para todo D_i e D_j diferindo em no máximo um elemento, ou seja, vizinhos.

Portanto, a sensibilidade irá mensurar o impacto de um indivíduo, ao ser removido do conjunto de dados, no resultado da consulta. A sensibilidade é de fundamental importância para determinar a quantidade de ruído adicionado na saída de um mecanismo, uma vez que quanto maior a sensibilidade, maior será a quantidade de ruído adicionada à saída do mecanismo para mascarar a ausência do indivíduo, de forma a garantir a sua privacidade (DOMINGO-FERRER *et al.*, 2016).

Os mecanismos mais comuns existentes na literatura para garantir a Privacidade Diferencial são o Exponencial e o de *Laplace* (DWORK *et al.*, 2014). O mecanismo Exponencial é aplicado sobre consultas não numéricas, enquanto que o de *Laplace*, que será empregado neste trabalho, atua sobre consultas numéricas de agregação, ou seja, consultas do tipo **COUNT**, **MIN**, **MAX**, **SUM** e **AVG**. Nesse mecanismo, o ruído a ser adicionado se baseia na geração de uma variável aleatória x, a qual segue uma distribuição de *Laplace* com média μ e escala b, através da fórmula:

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

Assim, o mecanismo de *Laplace* é definido formalmente por:

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

$$M = f(D) + Laplace(0, \frac{\Delta f}{\varepsilon})$$

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

2.2.2.1.4 Exemplo

Esta seção apresenta um exemplo explicativo acerca da Privacidade Diferencial, demonstrando o seu funcionamento em um conjunto de dados hipotético. Para isso, considere o conjunto de dados *D* representado pela Tabela 5, o qual pertence ao Departamento de Trânsito de algum estado brasileiro e contém o número de carros vinculados a cada indivíduo.

Tabela 5 – Exemplo de conjunto de dados original *D* contendo o número de carros de cada indivíduo.

ID	**Nome**	**Número de Carros**
1	Artur	3
2	Bernardo	1
3	Celso	2
4	Dener	6

Considere, também, que a consulta f a ser realizada sobre o conjunto de dados retorna a soma dos carros de todos os indivíduos. Temos que a resposta real da consulta é 12. No entanto, também é necessário aplicar a consulta sobre cada conjunto de dados vizinho do conjunto de dados original. A Tabela 6 apresenta os conjuntos de dados vizinhos e suas respectivas respostas para a mesma consulta.

Uma vez calculados todos os resultados da consulta depois de aplicá-la em cada conjunto de dados vizinho, é preciso calcular a variação máxima que a ausência de um indivíduo provoca no resultado da consulta. Isso é necessário para garantir a privacidade dos indivíduos, estando eles participando, ou não, do conjunto de dados. Como observado na Tabela 6, a variação máxima no resultado da consulta ocorre ao remover o indivíduo representado pelo registro de "ID" = 4, gerando o conjunto de dados vizinho mostrado na Tabela 6d. A sensibilidade da consulta é calculada através da maior diferença $|f(D) - f(D_i)|$, a qual ocorre com o valor de $i = 4$, gerando como resultado $|12 - 6| = 6$. Por fim, de forma a atender os requisitos do modelo de Privacidade Diferencial, a quantidade de ruído a ser adicionada ao resultado real da consulta, de forma a garantir a privacidade dos indivíduos, utilizando o mecanismo de *Laplace*, deve ser igual a $Laplace(0, \frac{6}{\varepsilon})$. O parâmetro ε, que necessita ser definido pelos *dataholders*, foi definido como 1 para o exemplo em questão.

Finalizando o exemplo, a Tabela 7 apresenta cinco possíveis valores de ruído, respos-

Tabela 6 – Conjuntos de dados vizinhos gerados a partir do conjunto de dados original D e suas respectivas respostas para a consulta f. (a) $f(D_1) = 9$. (b) $f(D_2) = 11$. (c) $f(D_3) = 10$. (d) $f(D_1) = 6$.

ID	Nome	Número de Carros
2	Bernardo	1
3	Celso	2
4	Dener	6

(a) Conjunto de dados D_1.

ID	Nome	Número de Carros
1	Artur	3
3	Celso	2
4	Dener	6

(b) Conjunto de dados D_2.

ID	Nome	Número de Carros
1	Artur	3
2	Bernardo	1
4	Dener	6

(c) Conjunto de dados D_3.

ID	Nome	Número de Carros
1	Artur	3
2	Bernardo	1
3	Celso	2

(d) Conjunto de dados D_4.

tas e suas respectivas probabilidades de ocorrência ao aplicar o mecanismo de *Laplace*. Como é possível observar, o valor de ruído de 7,28 possui a menor probabilidade de ocorrência (2,48%), enquanto o ruído de -1,61 possui a maior probabilidade (6,37%), resultando nas respostas anonimizadas de 19,28 e 10,39 carros, respectivamente. Devido às características da própria distribuição de *Laplace*, a probabilidade do mecanismo retornar uma resposta anonimizada próxima da resposta real da consulta (na consulta em questão, próximo do valor de 12 carros) será maior sempre que o ruído gerado estiver o mais próximo de *zero*. Contudo, todos os valores mostrados, resultantes do mecanismo de *Laplace*, satisfazem a Privacidade Diferencial.

Tabela 7 – Possíveis valores de ruído, respostas e probabilidades de ocorrência das respostas após a aplicação do mecanismo de *Laplace*.

Ruído	$f(D)$+ *ruído*	*Prob*($f(D)$+ *ruído*) (%)
-2,65	9,35	5,36
3,98	15,98	4,29
7,28	19,28	2,48
-1,61	10,39	6,37
-2,73	9,27	5,28

2.2.2.1.5 Desafios e Limitações

A Privacidade Diferencial foi proposta com o objetivo de prover informações estatísticas sobre um conjunto de dados e, dessa forma, garantir a privacidade dos indivíduos, uma vez que suas informações propriamente ditas não são disponibilizadas. No entanto, nada impede que um usuário malicioso com conhecimento prévio possa descobrir algo sobre um indivíduo, visto que esse mesmo indivíduo poderia nem fazer parte do conjunto de dados, mas poderia

ter semelhanças com os outros indivíduos nele presentes e, assim, o resultado proveniente do mecanismo poderia ser utilizado para descobrir informações sobre o indivíduo. Contudo, o modelo não considera isso uma violação de privacidade, visto que a Privacidade Diferencial é um modelo relativo, ou seja, garante que a participação, ou não, no conjunto de dados que será submetido a análises apenas aumenta ligeiramente o risco de descoberta (CLIFTON; TASSA, 2013).

Algumas características limitantes da Privacidade Diferencial envolvem o parâmetro ε, a sensibilidade da consulta e as saídas do mecanismo. O valor do parâmetro ε é difícil de ser estimado, visto que o seu valor não é uma medida direta da privacidade, mas um limitante no impacto que um indivíduo provoca no conjunto de dados. Por sua vez, a sensibilidade da consulta pode ser um tanto quanto complexa de ser gerada, além de ter um valor substancialmente alto em alguns casos, degradando completamente a utilidade das respostas geradas. Além disso, as resposta geradas pelo mecanismo podem, ainda, não ser muito adequadas para a utilização em algumas áreas específicas devido à sua maneira incerta de como são geradas, além de que, em alguns casos, as respostas podem não fazer muito sentindo, principalmente em consultas de contagem (**COUNT**), onde o mecanismo, na maioria das vezes, retorna números reais, o que soa um tanto quanto estranho por se tratar de uma consulta de contagem, a qual deveria ter um valor exato.

Por fim, a Privacidade Diferencial assume que os indivíduos presentes no conjunto de dados são independentes entre si, o que nem sempre é verdade. Em diversos cenários do mundo real, a presença de um determinado indivíduo no conjunto de dados pode causar inúmeras mudanças nas informações de outros indivíduos. Quando isso ocorre, dizemos que os dados são correlacionados. Entretanto, aplicar um mecanismo de Privacidade Diferencial em um conjunto de dados onde existe fortes evidências de existência de correlação entre os indivíduos, porém desconsiderando essa correlação, pode levar a severas violações de privacidade dos indivíduos. Surge, então, a Privacidade Diferencial Correlacionada, uma extensão da Privacidade Diferencial para cenários onde existe a correlação entre os indivíduos do conjunto de dados.

2.2.2.2 Privacidade Diferencial Correlacionada

A Privacidade Diferencial Correlacionada, proposta por (ZHU *et al.*, 2015), surgiu com o objetivo de suprir a limitação da Privacidade Diferencial no que diz respeito à independência dos indivíduos no conjuntos de dados, de forma que aplicar um mecanismo convencional

sobre um conjunto de dados que detém informações correlacionadas não é suficiente para garantir a privacidade dos indivíduos envolvidos.

O exemplo a seguir demonstra como as chances de um usuário malicioso descobrir informações acerca de um indivíduo aumentam ao desconsiderar a existência de correlação entre os indivíduos durante a utilização de um mecanismo de Privacidade Diferencial.

Suponha que um indivíduo, conhecido por "Hefesto", vive com mais 9 familiares em uma mesma casa. Suponha igualmente que uma nova doença, até então desconhecida, altamente contagiosa está se disseminando pelos arredores do local onde mora Hefesto e sua família. Um usuário malicioso que possui nenhum conhecimento sobre a saúde da família poderia executar a seguinte consulta: *"Quantas pessoas na família do Hefesto possuem a nova doença?"*. A resposta real para a consulta será *zero*, quando ninguém estiver infectado com a doença, ou 10, quando alguém tiver contraído a doença e infectado seus outros familiares. A sensibilidade em qualquer consulta de contagem nunca excederá o valor de 1, visto que, devido ao mecanismo considerar a independência entre os indivíduos, a remoção de qualquer um deles do conjunto de dados original provocará uma mudança de, no máximo, 1 em relação à resposta real da consulta. Assim, um mecanismo, ao executar $Laplace(\frac{1}{\varepsilon})$, adicionou ruído à resposta original da consulta e retornou o valor de 12 como resposta anonimizada. Uma reposta igual a 12 é $exp(10\varepsilon)$ vezes mais comum de ocorrer quando a resposta real é 10 (alguém está doente) em comparação a quando a resposta real é *zero* (ninguém está doente). Dessa forma, a correlação presente entre os indivíduos produziu um resultado que permitiu que a probabilidade de um usuário malicioso inferir algo acerca de um indivíduo fosse modificada em um fator igual a $exp(10\varepsilon)$, contrastando as garantias do modelo de Privacidade Diferencial, o qual afirma que a probabilidade de um usuário malicioso inferir algo não pode variar em um fator superior a $exp(\varepsilon)$.

Como visto no exemplo acima, diante da existência de correlação entre indivíduos de um conjunto de dados, as garantias do modelo de Privacidade Diferencial passam a não ser mais atendidas, colocando os indivíduos em situações de vulnerabilidade até mesmo contra usuários maliciosos com pouco, ou nenhum, conhecimento prévio. É preciso que a correlação entre os indivíduos seja previamente considerada, em contrapartida ao que foi proposto pelo modelo original, de forma a garantir a privacidade dos indivíduos. A inclusão do conceito de correlação no modelo de privacidade faz com que mudanças sejam necessárias em algumas definições apresentadas anteriormente, além de que outras precisem ser criadas, de forma a atender as garantias da Privacidade Diferencial sobre um conjunto de dados correlacionado.

Uma solução ingênua para garantir a privacidade dos indivíduos nesse caso seria multiplicar o valor da sensibilidade da consulta pelo número de indivíduos que atendem à consulta executada sobre o conjunto de dados. Dessa forma, a probabilidade de um usuário malicioso ser capaz de inferir algo sobre algum indivíduo seria drasticamente reduzida. Entretanto, a utilidade dos dados poderia ser totalmente destruída, uma vez que, dependendo da quantidade de indivíduos que atendem à consulta, o valor da sensibilidade poderia subir consideravelmente.

É necessário que a correlação entre os indivíduos seja mensurada cautelosamente e inserida de maneira razoável nas fórmulas existentes da Privacidade Diferencial. Dessa forma, a Privacidade Diferencial Correlacionada modifica o cálculo da sensibilidade ao considerar a correlação existente entre os indivíduos. Essa nova sensibilidade é chamada de sensibilidade correlacionada. No entanto, antes de defini-la, é preciso tomar conhecimento de alguns novos conceitos, começando pelo grau de correlação.

Definição 5 *Se dois indivíduos, r_i e r_j, são correlacionados entre si, o relacionamento é representado pelo Grau de Correlação $\delta_{ij} \in [-1,1]$ e $|\delta_{ij}| > \delta_0$, onde δ_0 é o limiar do grau de correlação.*

Visto que o grau de correlação representa apenas o relacionamento entre dois indivíduos, é possível listar todos os relacionamentos existentes entre os indivíduos por intermédio de uma matriz de graus de correlação, representada por:

Definição 6 *A Matriz de Graus de Correlação mantém os relacionamentos δ entre os os registros.*

$$\Delta = \begin{bmatrix} \delta_{11} & \delta_{12} & \delta_{13} & \dots & \delta_{1n} \\ \delta_{21} & \delta_{22} & \delta_{23} & \dots & \delta_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \delta_{n1} & \delta_{n2} & \delta_{n3} & \dots & \delta_{nn} \end{bmatrix}$$

Em resumo, se $\delta_{ij} < \delta_0$, o valor de δ_{ij} é definido como *zero*. Além disso, se $\delta_{ij} = 0$, não existe nenhuma correlação entre os registros r_i e r_j. Opostamente, se $\delta_{ij} = 1$ ou -1, os registros r_i e r_j são completamente correlacionados. Quanto mais próximo de -1 ou 1 o valor de δ_{ij} for, mais forte será a correlação entre os registros r_i e r_j e maior será o impacto sobre o registro r_j ao remover o registro r_i do conjunto de dados. Assim, o grau de correlação mostra o impacto que um registro tem sobre outro.

Uma vez definido, é preciso propor meios para identificar e mensurar o parâmetro δ_{ij}. Em outras palavras, é necessário realizar um estudo da correlação dos dados. Um dos possíveis meios aplicados para mensurar a correlação entre indivíduos consiste em utilizar informações adquiridas previamente. Entretanto, esse tipo de análise requer que o responsável por essa tarefa tenha plena *expertise* sobre o domínio dos dados. Identificar os indivíduos fortemente correlacionados pode ser uma tarefa relativamente fácil, enquanto que identificar relacionamentos com um baixo grau de correlação pode se tornar bastante desafiador.

Diante das dificuldades expostas para identificar os relacionamentos entre os indivíduos, surgem duas estratégias para calcular a correlação entre indivíduos de forma automática, ou seja, sem a necessidade de conhecimentos adquiridos por terceiros para estimar o grau de correlação. Tais estratégias consistem no Coeficiente de Correlação de *Pearson* (PEARSON, 1895) e no Coeficiente de Correlação de Postos de *Spearman* (SPEARMAN, 1904). Ambos são amplamente utilizados na área da estatística e consistem em métricas que avaliam a dependência entre duas variáveis, ou seja, a intensidade da relação entre duas variáveis. Dessa forma, tanto o coeficiente de *Pearson* quanto o de *Spearman* podem ser perfeitamente aplicados no contexto de dados correlacionados para identificar os relacionamentos entre indivíduos e quantificar o quão semelhantes são entre si. Indivíduos mais semelhantes possuem um maior coeficiente de correlação, e vice-versa.

Uma vez de posse dos graus de correlação devidamente calculados, é possível definir a sensibilidade do registro, a qual representa o impacto de um registro r_i, ao removê-lo do conjunto de dados D, no resultado de uma consulta f. A sensibilidade do registro é de extrema importância, além de servir como base, para calcular a sensibilidade correlacionada.

Definição 7 *Dado um Δ, uma consulta f e um conjunto de dados D, a Sensibilidade do Registro r_i é:*

$$CS_i = \sum_{j=0}^{n} |\delta_{ij}|(||f(D^j) - f(D^{-j})||_1)$$

para todo D^j e D^{-j} diferindo em um elemento j.

Por fim, a sensibilidade correlacionada irá mensurar o impacto de um indivíduo, ao ser removido do conjunto de dados correlacionado, no resultado da consulta.

Definição 8 *Para uma consulta f, a Sensibilidade Correlacionada lista todos os registros r que compõem a resposta de f e seleciona o maior valor da Sensibilidade do Registro como sendo a Sensibilidade Correlacionada,*

$$CS_f = \max_{i \in f}(CS_i)$$

Assim, as respostas das consultas serão dadas a partir da forma:

$$M = f(D) + Laplace(\frac{CS_f}{\varepsilon})$$

2.3 Agrupamento de Dados

2.3.1 Conceitos Básicos

Agrupamento de dados, ou *clustering*, é o nome dado para o grupo de técnicas computacionais que consiste em separar objetos em grupos, baseando-se nas características que esses objetos têm em comum. A ideia básica consiste em colocar, em um mesmo grupo, os objetos mais similares entre si baseando-se em algum critério pré-determinado, de forma que elementos de um mesmo grupo são mais similares entre si do que entre os elementos de outros grupos. O critério baseia-se, normalmente, em uma função de similaridade, ou dissimilaridade, que pode variar de acordo com o algoritmo escolhido ou de acordo com os dados utilizados. Uma vez definida, a função recebe dois objetos como entrada e retorna a distância entre eles (LINDEN, 2009).

Os objetos também podem ser chamados de elementos, exemplos, tuplas ou registros. Cada objeto representa uma entrada de dados que pode ser constituída por um vetor de atributos. Um atributo pode ser numérico ou categórico. Exemplos de atributos numéricos, representados por números inteiros ou reais, são: idade, peso, altura, salário, entre outros. Já alguns exemplos de atributos categóricos são: tipo sanguíneo, preferência sexual, patente militar, além de atributos representados por valores booleanos, como atributos que informam se a pessoa está doente, ou não (GORDON, 1999).

As técnicas de agrupamento de dados são ferramentas muito úteis para a análise de dados em diversas situações distintas. No contexto de dados correlacionados, técnicas de agrupamento podem ser utilizadas para identificar os relacionamentos entre indivíduos. Ou

seja, os indivíduos pertencentes a um mesmo grupo se relacionam com os demais indivíduos do mesmo grupo. No entanto, uma vez que os algoritmos de agrupamento são técnicas não supervisionadas, não se sabendo a priori quais e quantos grupos (*clusters*) serão encontrados, encontrar uma configuração de agrupamento ótima pode ser uma tarefa bastante complexa, visto que apresenta uma complexidade exponencial.

Os principais métodos de agrupamento existentes na literatura são classificados em hierárquicos ou baseados em particionamento, densidade, grade ou modelos de mistura. No entanto, o foco deste trabalho será em dois métodos específicos, o *DBSCAN* (ESTER *et al.*, 1996) e o *GMM* (BISHOP, 2006). O primeiro baseia-se em densidade, enquanto o segundo em modelos de mistura.

Em métodos baseados em densidade, os *clusters* são modelados como regiões densas do conjunto de dados, divididas por áreas de regiões esparsas. Os *clusters* são compostos pelos conjuntos de maior número de conexões próximas. Dentre as vantagens desses métodos pode-se destacar a capacidade de identificar *clusters* de formatos arbitrários, enquanto que outros tipos de métodos existentes, como os baseados em particionamento, apresentam melhores resultados em *clusters* de formatos circulares.

Técnicas de agrupamento baseadas em modelos de mistura assumem que os dados são gerados a partir de uma composição de distribuições, de forma que, a partir dos dados, uma técnica busca recuperar o modelo original. Depois de construído o modelo, é possível definir os *cluster* a partir dele. Dentre as vantagens desse tipo de técnica, além de serem capazes de identificar *clusters* de formatos arbitrários, possuem uma maior flexibilidade no processo de identificação dos *clusters*, visto que são capazes de se adaptar a diversas distribuições de dados, sejam elas de Bernoulli, Gaussiana ou qualquer distribuição pertencente a outra família.

2.3.2 DBSCAN

O *DBSCAN* é um algoritmo de agrupamento baseado em densidade amplamente utilizado pela comunidade científica. Seu principal objetivo consiste em identificar concentrações de objetos que estão espacialmente próximos um dos outros. Em outras palavras, o algoritmo busca por pontos (elementos) que possuem mais que um certo limiar de vizinhos em um determinado raio. Dessa forma, caso um elemento p satisfaça essa propriedade, os vizinhos de p também pertencerão ao mesmo cluster de p e o mesmo processo será aplicado sobre todos os seus vizinhos.

Outras vantagens do *DBSCAN*, além da capacidade de encontrar *clusters* de formatos arbitrários, consistem na facilidade de configurá-lo, devido ao seu reduzido número de parâmetros de entrada, e sua robustez contra *outliers*, i.e., elementos que não pertencem a nenhum *cluster*, sendo, a última, uma característica não muito presente na maioria dos algoritmos de agrupamento. Com respeito à sua configuração, além do conjunto de dados a ser agrupado, o *DBSCAN* recebe também dois parâmetros de entrada: *eps* e *minPoints*. O primeiro parâmetro refere-se ao raio no qual a verificação de vizinhança será realizada. Este parâmetro vai capturar a função de similaridade encontrada em outras estratégias de agrupamento. Já o segundo refere-se à quantidade mínima de elementos em um certo raio de vizinhança para a formação de um *cluster*. A função de distância utilizada para determinar a vizinhança de um determinado ponto deve ser definida baseando-se no tipo de dado a ser agrupado. É importante ressaltar que a função de distância deve obedecer à algumas restrições típicas de qualquer função de distância, como: simetria e desigualdade triangular. Além disso, a distância entre quaisquer dois pontos p e q somente será igual a *zero* se $p = q$.

A Figura 4 apresenta um exemplo de um conjunto de dados com os agrupamentos encontrados pelo algoritmo *DBSCAN*, onde seis *clusters* foram identificados e cada um deles representado em uma cor distinta, enquanto que os *outliers*, os pontos que não fazem parte de qualquer agrupamento, são representados na cor azul mais escura.

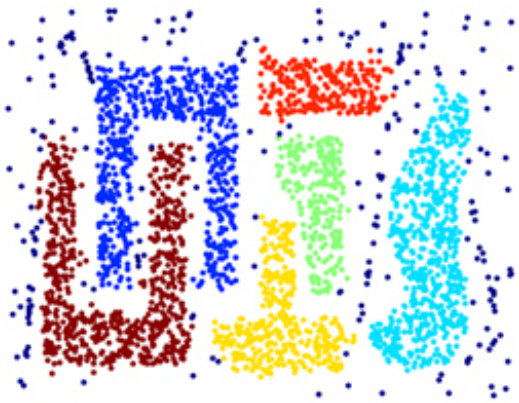


Figura 4 – *Clusters* de formatos arbitrários encontrados pelo algoritmo *DBSCAN* (THE..., 2017).

Algumas outras definições básicas, mas não menos importante, utilizadas no algoritmo *DBSCAN* são apresentadas a seguir:

- $|A|$: cardinalidade do conjunto A.
- $N_{eps}(p)$: conjunto de pontos q que estão a uma distância menor que eps em relação ao

ponto p. Também chamado de conjunto dos vizinhos de p.

- *Directly Density-Reachable* / Diretamente Alcançável por Densidade (*DDR*): um ponto p é dito *DDR* a partir de um ponto q se $p \in N_{eps}(q)$ e $|N_{eps}(q)| \geq minPoints$.
- *Density-Reachable* / Alcançável por Densidade (*DR*): um ponto p é dito *DR* a partir de um ponto q se existe uma sequência de pontos $p_1, ..., p_n$ onde $p_1 = p$ e $p_n = q$, tal que p_{i+1} é *DDR* a partir de p_i.
- *Density-Connected* / Conectado por Densidade (*DC*): um ponto p está conectado por densidade a um ponto q se existe um ponto o tal que p e q são *DR* a partir de o.
- Ponto central (*Core point*): um ponto p é classificado como ponto central se $N_{eps}(p) \geq minPoints$.
- Ponto de fronteira (*Border point*): um ponto p é classificado como ponto de fronteira se $N_{eps}(p) < minPoints$ e p é *DDR* a partir de um ponto central.
- Ruído (*Noise*): um ponto p é classificado como ruído se $|N_{eps}(p)| < minPoints$ e p não é *DDR* a partir de nenhum ponto central.

Por fim, um *cluster C* é definido como um subconjunto não vazio que satisfaz as propriedades de maximalidade e conectividade, definidas na forma:

- Maximalidade: para quaisquer dois pontos p e q, se $p \in C$ e q é *DR* a partir de p, então $q \in C$.
- Conectividade: para quaisquer dois pontos p e $q \in C$, p e q são *DC*.

Apesar da série de vantagens que o método dispõe, algumas limitações ainda são evidentes. Uma dessas limitações consiste em determinar os parâmetros de entrada do algoritmo, eps e $minPoints$, visto que influenciam diretamente no resultado do agrupamento. Determinar os valores ideais para esses parâmetros pode se tornar uma tarefa bastante complexa, exigindo que o responsável por essa tarefa tenha um conhecimento prévio sobre os dados a serem agrupados. Ainda assim, em sua grande maioria, os parâmetros são definidos empiricamente. Sabe-se que valores de eps muito baixos tendem a gerar *clusters* muito esparsos e com muitos *outliers*, enquanto valores muito altos geram *clusters* muito densos. Portanto, é necessário encontrar valores adequados para os parâmetros, de forma a identificar *clusters* condizentes com a disposição dos dados e gerar resultados satisfatórios. Além disso, cada objeto do conjunto de dados pertence, exclusivamente, a um único *cluster*. Em outras palavras, dado um objeto com características similares a dois grupos distintos, ele terá que ser alocado a apenas um dos dois *clusters*. Isso acaba se tornando uma grande desvantagem ao se aplicar o *DBSCAN*

em um conjunto de dados com a presença de indivíduos correlacionados, visto que uma vasta quantidade de possíveis relacionamentos existentes entre indivíduos de *clusters* distintos serão desconsiderados.

2.3.3 GMM

O *GMM* é um algoritmo de agrupamento baseado em modelos de mistura que pressupõe que os dados a serem agrupados são gerados a partir de uma distribuição de mistura de Gaussianas composta por várias componentes (*clusters*). Modelos de mistura são famílias de distribuições formadas pela composição de mais de uma distribuição. Esses modelos são formados pela distribuição de probabilidade P_1, com probabilidade w_1, pela distribuição de probabilidade P_2, com probabilidade w_2, e assim por diante. Dessa forma, a densidade resultante é a combinação ponderada das densidades que caracterizam essas distribuições, onde as respectivas probabilidades são os fatores de ponderação. As distribuições $P_1, ..., P_k$ são chamadas de componentes da mistura, enquanto que as probabilidades $w_1, ..., w_k$ são chamadas de pesos da mistura.

Definição 9 *Dado um conjunto de dados $X = x_1, ..., x_N$ de dimensão dim e composto por N elementos, um Modelo de Mistura de Gaussianas é definido através da função densidade de probabilidade $p(X|\lambda)$ dada por:*

$$p(X|\lambda) = \sum_{k=1}^{K} w_k g(X|\theta_k),$$

onde K é o número de distribuições que formam a mistura, θ_k corresponde ao conjunto de parâmetros definidos pela k-ésima componente da mistura, média e matriz de covariância, representadas pelo vetor de parâmetros $\theta_k = (\mu_k, \Sigma_k)$, e w_k são as probabilidades, também chamadas de pesos da mistura, onde $w_k \geq 0$ e $\sum_{k=1}^{K} w_k = 1$. O vetor $\lambda = w_1, ..., , w_K, \theta_1, ..., \theta_K$ representa o conjunto de parâmetros da mistura. Cada componente $g(X|\theta_k)$ da mistura é uma distribuição Gaussiana dim-variada representada pela função de densidade de probabilidade definida por:

$$g(X|\theta_k) = \frac{1}{(2w)^{\frac{dim}{2}} |\Sigma_k|^{\frac{1}{2}}} e^{-\frac{1}{2}(X-\mu_k)^T \Sigma_k^{-1} (X-\mu_k)}.$$

No *GMM*, a medida de similaridade é a probabilidade *a posteriori* dos dados em relação às classes. Seu objetivo é maximizar o critério que avalia a qualidade do agrupamento, a função de verossimilhança. Além de uma técnica de agrupamento de dados, o *GMM* fornece uma descrição estatística sobre os dados através dos parâmetros da mistura: média (μ) e matriz de covariância (Σ).

Em modelos de mistura, gerados a partir de uma distribuição de mistura composta por k componentes, é pressuposto que cada elemento do conjunto de dados a ser agrupado é gerado a partir de uma componente e que elementos de grupos distintos são gerados a partir de componentes diferentes. Porém, cada elemento possui uma certa probabilidade de pertencer a qualquer um dos *clusters*, onde aquele que possui a maior probabilidade é dito como o *cluster* a que um determinado elemento pertence. A Figura 5 apresenta um exemplo de um conjunto de dados com os agrupamentos encontrados pelo algoritmo *GMM*, onde 4 *clusters*, representados por Gaussianas, foram identificados e cada um deles representado por um símbolo de cor distinta no centro de cada Gaussiana. Embora os pontos que não se encontram em nenhum *cluster* possam ser confundidos com *outliers*, todos os pontos possuem, por mais que baixa, uma certa probabilidade de pertencerem a qualquer um dos *clusters* identificados.

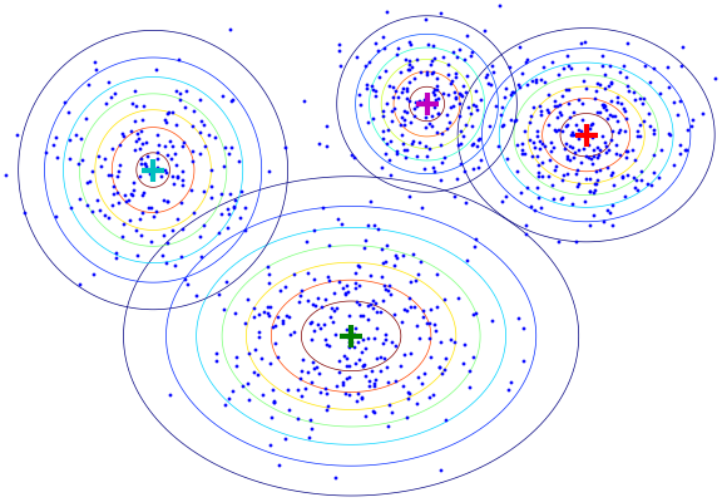


Figura 5 – *Clusters* encontrados pelo algoritmo *GMM* (ZHU, 2014).

A etapa de formação dos *clusters* consiste em identificar e recuperar um modelo de mistura a partir do conjunto de dados a ser agrupado. Acontece que, inicialmente, não se tem nenhum conhecimento acerca de qual componente gerou cada elemento do conjunto de dados, dos parâmetros (λ) das componentes e do número de componentes K. Uma solução para estimar os valores dos pesos (w) e parâmetros (θ) que caracterizam as distribuições, até então

desconhecidos, consiste em utilizar a técnica de *Maximum Likelihood Estimation* / Máxima Verossimilhança (*MLE*) (SCHOLZ, 1985) para maximizar o logaritmo da seguinte função de verossimilhança:

$$\begin{aligned}\log p(X|\lambda) &= \log\left(\prod_{n=1}^{N} p(x_n|\lambda)\right) \\ &= \sum_{n=1}^{N} \log \sum_{k=1}^{K} w_k g(x_n|\theta_k).\end{aligned}$$

A função $p(X|\lambda)$ é chamada de função de verossimilhança. Visto que o logaritmo da verossimilhança é crescente, o conjunto de parâmetros λ que maximiza o logaritmo da verossimilhança também maximiza a verossimilhança. O objeto é encontrar uma boa estimativa de λ através da maximização do logaritmo de verossimilhança utilizando apenas os conjunto de dados de entrada X. Assim, quanto melhor a estimativa de λ, maior será a probabilidade de se obter os dados observados. O algoritmo *Expectation Maximization* / Maximização de Expectativa (*EM*) é bastante conhecido na literatura e amplamente utilizado para estimar os parâmetros que maximizam a verossimilhança em um modelo de mistura de Gaussianas.

Por sua vez, o número de componentes, ou *clusters*, pode ser estimado através da técnica conhecida como *Bayesian Information Criterion* / Critério de Informação Bayesiano (*BIC*). Resumidamente, a técnica consiste em identificar um número ótimo de componentes que se adeque bem ao conjunto de dados, baseando-se em uma função de verossimilhança, de forma que o subajuste (*underfitting*) e o sobreajuste (*overfitting*) sejam evitados. Mais detalhes serão apresentados na Seção 2.3.3.2

2.3.3.1 Maximização de Expectativa

Proposto por (DEMPSTER *et al.*, 1977), o algoritmo iterativo *EM* consiste em uma solução para os problemas de estimativa de parâmetros, comumente encontrados em modelos de mistura. A principal ideia por trás do método consiste em estimar os parâmetros desejados através de duas etapas alternadas, denominadas: etapa E (*E-Step*), também chamada de Expectativa (*Expectation*), e etapa M (*M-Step)*, ou Maximização (*Maximization*). As duas etapas são repetidas de maneira alternada até alcançar a convergência do *GMM*.

2.3.3.1.1 Etapa E

Nessa etapa, calcula-se a probabilidade de cada elemento pertencer a cada uma das componentes. Além disso, uma nova estimativa da função de verossimilhança é calculada através da seguinte equação:

$$\gamma(w_k, \theta_k|x_n) = \frac{w_k g(x_n|\theta_k)}{\sum_{k=1}^{K} w_k g(x_n|\theta_k)}.$$

Assim, a função γ representa a probabilidade do elemento x_n pertencer à k-ésima componente, onde $1 \leq j \leq K$.

2.3.3.1.2 Etapa M

Por sua vez, a etapa M consiste em maximizar as componentes da mistura, atualizando seus parâmetros, através das equações:

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(w_k, \theta_k|x_n) x_n$$
$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(w_k, \theta_k|x_n)(x_n - \mu_k)(x_n - \mu_k)^T$$
$$w_k = \frac{N_k}{N},$$

onde N_k representa o número de elementos que pertencem ao *cluster k*. Em outras palavras, visto que cada elemento possui uma certa probabilidade de pertencer aos *K clusters*, a variável N_k consiste no número de elementos que possuem maior probabilidade de pertencer ao *cluster k*.

O exemplo da Figura 6 apresenta o funcionamento do algoritmo *EM* considerando um número máximo de 20 iterações até atingir a convergência do *GMM*. Inicialmente, destaca-se em (a), na cor verde, os pontos que representam o conjunto de dados. Em (b), os *clusters* são estimados pela primeira vez e representados pelas cores azul e vermelho. De (c)-(e) os *clusters* são recalculados e aperfeiçoados ao longo das iterações, representada por L. Por fim, o algoritmo *GMM* atinge a convergência após as 20 iterações e os *clusters* resultantes são representados em (f).

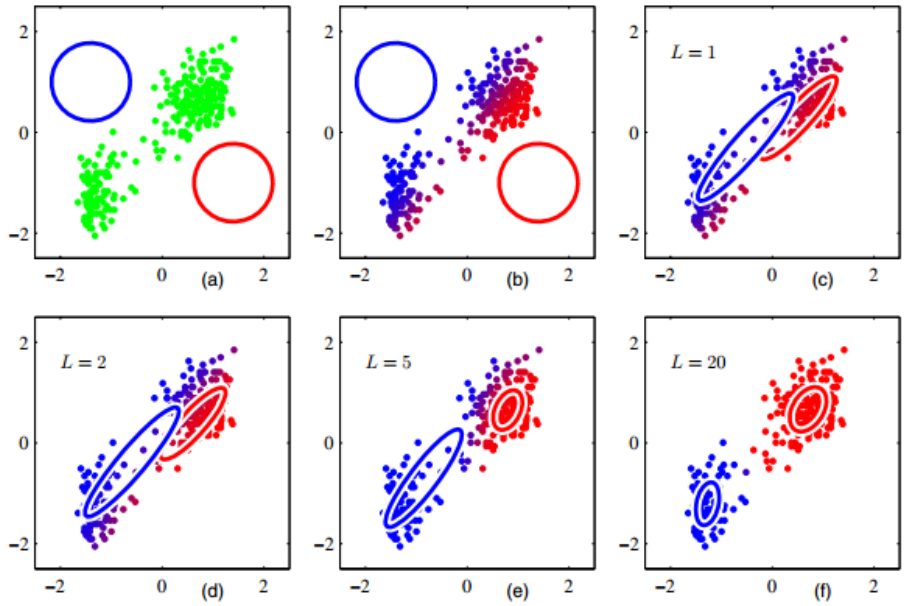


Figura 6 – Funcionamento do algoritmo *EM* após 20 iterações (BISHOP, 2006).

2.3.3.2 Critério de Informação Bayesiano

Primeiramente, antes de entrarmos nos detalhes do *BIC*, é preciso saber que não existem modelos verdadeiros, apenas modelo aproximados que geram perda de informação. Em outras palavras, é necessário encontrar o melhor modelo, dentre todos os que foram ajustados, para explicar o fenômeno que está sendo estudado, em nosso caso, o modelo que melhor representa um conjunto de dados. Esse é um dos motivos os quais tornam a utilização da *EM*, recém discutida, e do *BIC* essenciais durante o processo de encontrar o modelo mais próximo do ideal.

Durante o processo de ajuste, é possível melhorar o modelo através da maximização da função de verossimilhança por meio de seus parâmetros. No entanto, isso pode levar o modelo ao sobreajuste, que ocorre quando o modelo se ajusta demasiadamente ao conjunto de dados. Para evitar que isso venha a ocorrer, o *BIC* introduz uma penalidade no modelo de acordo com o seu número de parâmetros. Isso garante que o modelo encontrado por meio do *BIC* seja uma boa representação, não viciada, do conjunto de dados.

O *BIC* foi proposto por (SCHWARZ *et al.*, 1978) e é dado por:

$$BIC = -2\log p(X|\hat{\theta}) + K\log N,$$

onde $p(X|\hat{\theta})$ é a função de verossimilhança do modelo, K é o número de parâmetros a serem estimados e N é o número de elementos do conjunto de dados X.

Cada modelo distinto gerado terá o seu *BIC*. Assim, caso desejemos encontrar o modelo que melhor representa um conjunto de dados hipotético, precisaremos construir diversos modelos variando seus números de componentes e, para cada modelo gerado, calcular o *BIC*. Ao final, o modelo com o menor *BIC* deve ser escolhido.

2.4 Conclusão

Neste capítulo apresentamos conceitos e definições sobre preservação de privacidade de dados. Foram apresentados modelos de privacidade sintáticos e interativos, dando ênfase aos modelos interativos e, em especial, ao modelo de Privacidade Diferencial, o qual é empregado na solução deste trabalho. Além disso, algumas limitações do modelo de Privacidade Diferencial foram mencionadas, principalmente quando aplicado sobre contextos onde existem relacionamentos entre os indivíduos, de maneira que o modelo necessita ser adaptado a fim de continuar garantindo a privacidade dos indivíduos. Por fim, apresentamos duas técnicas de agrupamento, o *DBSCAN* e o *GMM*, destacando seus benefícios e como poderiam ser aplicados sobre um contexto de dados correlacionados com o objetivo de identificar os indivíduos que se relacionam entre si.

3 TRABALHOS RELACIONADOS

Neste capítulo apresentaremos alguns trabalhos relacionados à preservação de privacidade de indivíduos no contexto de dados correlacionados utilizando o modelo de Privacidade Diferencial. Além disso, também apresentaremos alguns trabalhos que empregam técnicas de agrupamento em suas respectivas soluções para o problema da preservação de privacidade, mesmo que o contexto desses trabalhos são seja a correlação de dados. Como visto anteriormente, no Capítulo 2, se não for dada a devida atenção à correlação de dados ao se utilizar um mecanismo diferencialmente privado, um usuário malicioso pode se tornar capaz de inferir informações acerca dos indivíduos, até mesmo daqueles que sequer estão presentes no conjunto de dados.

Uma vez que o modelo de Privacidade Diferencial pode ser empregado tanto em ambientes interativos quanto em ambiente não-interativos, i.e., para a publicação de dados, optamos por classificar os trabalhos relacionados em estratégias voltadas para dados correlacionados em geral, Seção 3.1, e estratégias que empregam técnicas de agrupamento de dados, Seção 3.2, visto que foi uma das estratégias utilizadas em nossa solução para identificar os relacionamentos existentes entre os indivíduos de um conjunto de dados.

3.1 Privacidade Diferencial e Dados Correlacionados

O modelo de Privacidade Diferencial foi proposto, inicialmente, para proteger a privacidade dos indivíduos assumindo que os dados são independentes. No entanto, visto que em muitos contextos, principalmente em redes sociais onde muitos indivíduos se relacionam entre si, a Privacidade Diferencial passou a não mais ser garantida devido à existência desses relacionamentos entre os indivíduos. Portanto, é necessário adaptar o modelo, identificando os relacionamentos, para continuar garantindo a privacidade dos indivíduos. Diversos trabalhos (KIFER; MACHANAVAJJHALA, 2014; CHEN *et al.*, 2014; ZHU *et al.*, 2015; LIU *et al.*, 2016) foram propostos com o objetivo de adaptar o modelo de Privacidade Diferencial sobre dados correlacionados, i.e. dados onde existem relacionamentos entre indivíduos. Entretanto, alguns deles são impraticáveis, comprometendo a utilidade dos dados ou não provendo uma solução viável para identificar os relacionamentos existentes entre os indivíduos.

3.1.1 Estratégia proposta em (KIFER; MACHANAVAJJHALA, 2014)

Um dos pioneiros a tratar a utilização do modelo de Privacidade Diferencial em conjuntos de dados correlacionados, os autores constataram, em seu primeiro trabalho (KIFER; MACHANAVAJJHALA, 2011), que caso seja evidenciada a correlação entre os indivíduos e a mesma seja ignorada, os dados resultantes, sejam eles por meio de consultas ou publicação, proverão menos garantias de privacidade que o esperado.

Visando suprir essa limitação expressa em cenários os quais evidenciam relacionamentos entre os indivíduos e atender as garantias do modelo de Privacidade Diferencial, esse trabalho propõe a criação de um *framework* de privacidade customizável, denominado *Pufferfish*.

O *framework* pode ser utilizado para criar novas definições de privacidade, as quais são customizáveis de acordo com as necessidades de cada aplicação. Para isso, é necessário definir uma série de requisitos, denominados: segredos potenciais (*pontential secrets*), pares discriminativos (*discriminative pairs*) e cenários de evolução (*evolution scenarios*). Os segredos potenciais consistem em uma especificação explícita do que precisa ser protegido. Exemplos de segredos potenciais são: "o registro do indivíduo h_i está no conjunto de dados", "o registro do indivíduo h_i não está no conjunto de dados" ou "o volume de consultas é da ordem de 1-5 milhões de consultas". Por sua vez, os pares discriminativos são definições de como proteger os segredos potenciais, de forma que um usuário malicioso não seja capaz de distinguir o que é verdade, ou não. São exemplos de pares discriminativos: ("João está na tabela", "João não está na tabela") ou ("Pedro está doente", "Pedro não está doente"). Por fim, os cenários de evolução podem ser vistos como um conjunto de hipóteses sobre, por exemplo, como os dados evoluíram ou sobre o conhecimento adquirido por potenciais usuários maliciosos.

Uma vez definidos os conjuntos de requisitos, é possível utilizar algum mecanismo de privacidade que satisfaça as novas definições de privacidade, resultantes a partir dos requisitos, e assegurar a privacidade dos indivíduos, mesmo existindo correlação entre os mesmos. Além disso, o *Pufferfish* possibilita que os requisitos mencionados anteriormente sejam definidos por especialistas sem conhecimento em privacidade, sendo exigido apenas o conhecimento sobre o domínio da aplicação.

3.1.2 Estratégia proposta em (CHEN et al., 2014)

Embora também seja voltado para o contexto de dados correlacionados utilizando o modelo de Privacidade Diferencial, o trabalho se difere um pouco dos demais ao optar por focar, especificamente, em conjuntos de dados de redes sociais. A solução empregada é não-interativa, de forma que o conjunto de dados é modificado através do mecanismo Exponencial antes de ser publicado para os devidos fins de análise. Além disso, os autores afirmam que se trata do primeiro trabalho a propor uma solução prática para a publicação de conjuntos de dados de redes sociais por meio de um mecanismo de Privacidade Diferencial.

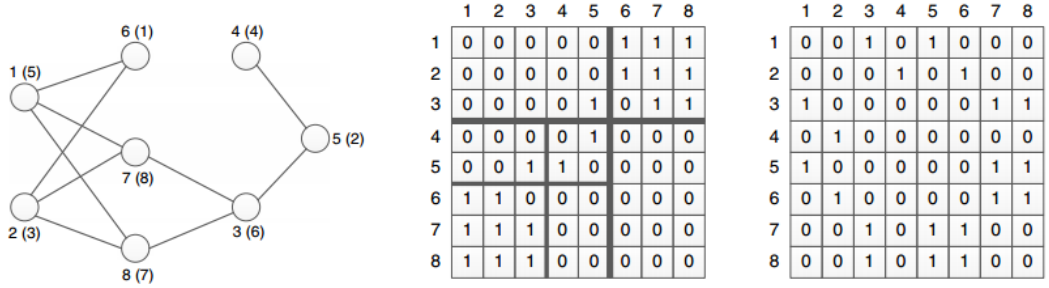
Conjuntos de dados de redes sociais são, normalmente, representados por meios de grafos, na forma: $G = (V, E)$, onde G representa o grafo, V os vértices, ou nós, que são os indivíduos, e E as arestas, que são os relacionamentos entre os indivíduos. No contexto de Privacidade Diferencial para grafos já se evidenciou dois possíveis tipos de ataques envolvendo os vértices e arestas: a re-identificação dos nós (*node re-identification*) e a divulgação de aresta (*edge disclosure*). O primeiro ataque objetiva identificar a quem um determinado nó pertencente, enquanto que o segundo é voltado para a descoberta do relacionamento entre os indivíduos, sendo o objetivo desse trabalho impedir esse tipo de ataque.

Uma vez que o modelo de Privacidade Diferencial originalmente proposto é vulnerável à correlação dos dados, não provendo as devidas garantias de privacidade, a solução descrita nesse trabalho consiste em aperfeiçoar o modelo introduzindo um parâmetro extra, denominado k, com o intuito de mensurar a extensão da correlação, levando-a em consideração para uma publicação dos dados que não comprometa a privacidade dos indivíduos envolvidos. O parâmetro k representa o número máximo de arestas conectadas aos vértices. Em outras palavras, cada vértice não pode possuir mais do que k conexões com os demais vértices.

O trabalho é divido em três etapas, são elas: Identificar Rótulos dos Vértices (*Identify Vertex Labeling*), Explorar Regiões Densas (*Explore Dense Region*) e Planejar Arestas (*Arrange Edge*). Antes de iniciar a primeira etapa, o valor do parâmetro k é utilizado para dividir o valor do parâmetro ε do modelo de Privacidade Diferencial, onde o valor resultante é dividido em três porções, sendo cada uma delas utilizadas em cada uma das etapas seguintes.

Na primeira etapa, busca-se identificar uma rotulação ótima dos vértices do grafo. Através dos rótulos é possível construir uma matriz de adjacência correspondente ao grafo, de forma que a matriz formada possua regiões densas, representadas por 1's. A rotulação dos vértices é realizada através de um algoritmo guloso que permuta os valores dos rótulos entre os

vértices a fim de obter um melhor contraste de densidade. A Figura 7 apresenta um conjunto hipotético de dados, representado por meio de um grafo, e duas possíveis matrizes de adjacência, a primeira composta pelos rótulos fora dos parênteses e a segunda pelos rótulos entre parênteses. Observe que o processo de rotulação tem impacto direto sobre a matriz de adjacência resultante e, caso não realizado de maneira adequada, pode acabar gerando matrizes que não evidenciam as regiões densas existentes no conjunto de dados.



	1	2	3	4	5	6	7	8
1	0	0	0	0	0	1	1	1
2	0	0	0	0	0	1	1	1
3	0	0	0	0	1	0	1	1
4	0	0	0	0	1	0	0	0
5	0	0	1	1	0	0	0	0
6	1	1	0	0	0	0	0	0
7	1	1	1	0	0	0	0	0
8	1	1	1	0	0	0	0	0

	1	2	3	4	5	6	7	8
1	0	0	1	0	1	0	0	0
2	0	0	0	1	0	1	0	0
3	1	0	0	0	0	0	1	1
4	0	1	0	0	0	0	0	0
5	1	0	0	0	0	0	1	1
6	0	1	0	0	0	0	1	1
7	0	0	1	0	1	1	0	0
8	0	0	1	0	1	1	0	0

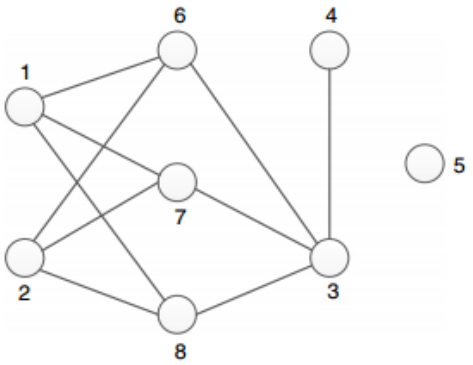
Figura 7 – Conjunto de dados hipotético e duas possíveis representações de matrizes de adjacência (CHEN *et al.*, 2014).

Na segunda etapa, a matriz de adjacência construída na etapa anterior é utilizada em um processo diferencialmente privado, por meio do mecanismo exponencial, combinado a um particionamento de dados, o qual adapta a estrutura de dados *quadtree* para explorar as regiões densas da matriz de adjacência. O processo resulta em uma *quadtree* acrescida de ruído, de forma que os nós da árvore representam as regiões densas da matriz e estes são associados a valores acrescidos de ruído. A partir dessa *quadtree*, é possível construir uma versão anonimizada da matriz de adjacência com boa acurácia.

Por fim, na terceira etapa, a *quadtree* obtida anteriormente é utilizada em um processo de reconstrução da matriz de adjacência e, consequentemente, do grafo. O processo é realizado por meio de um algoritmo que busca maximizar as métricas de utilidade. A Figura 8 apresenta a versão anonimizada do grafo ilustrado anteriormente, na Figura 7, e sua respectiva matriz de adjacência, também anonimizada, após a conclusão das etapas duas últimas etapas.

3.1.3 Estratégia proposta em (ZHU et al., 2015)

Este trabalho apresenta um novo modelo de privacidade, denominado Privacidade Diferencial Correlacionada, o qual propõe mudanças na análise dos relacionamentos entre indivíduos. Em seu estudo, os autores afirmam que os trabalhos (KIFER; MACHANAVAJJHALA, 2014) e (CHEN *et al.*, 2014) não tratam suficientemente bem a existência da correlação nos dados.



	1	2	3	4	5	6	7	8
1	0	0	0	0	0	1	1	1
2	0	0	0	0	0	1	1	1
3	0	0	0	1	0	1	1	1
4	0	0	1	0	0	0	0	0
5	0	0	0	0	0	0	0	0
6	1	1	1	0	0	0	0	0
7	1	1	1	0	0	0	0	0
8	1	1	1	0	0	0	0	0

Figura 8 – Versão anonimizada do grafo ilustrado na Figura 7 e sua respectiva matriz de adjacência, também anonimizada (CHEN *et al.*, 2014).

Segundo os autores, o primeiro trabalho não garante as propriedades do modelo de Privacidade Diferencial, enquanto que o método empregado no segundo trabalho adiciona uma quantidade de ruído excessiva às respostas reais das consultas que, embora garanta as propriedades exigidas pelo modelo, distorce demasiadamente a utilidade dos dados.

Nesse trabalho, os autores propõem mudanças no cálculo da sensibilidade das consultas através da inserção de um fator extra, denominado grau de correlação. Assim, a sensibilidade passa a se chamar sensibilidade correlacionada. O grau de correlação mede o quão similar dois indivíduos são entre si, e o seu valor impacta diretamente no valor da sensibilidade. O objetivo da sensibilidade correlacionada é levar em consideração os relacionamentos existentes entre os indivíduos durante o seu cálculo, de maneira a garantir a privacidade dos indivíduos.

Assim como no modelo de Privacidade Diferencial (DWORK, 2006), o qual identifica o indivíduo que causa o maior impacto sobre o resultado da consulta para proteger a identidade dos outros indivíduos, o modelo de Privacidade Diferencial Correlacionada parte do mesmo pressuposto, diferindo apenas pela existência do grau de correlação que, ao levar em consideração a existência de relacionamentos entre os indivíduos, resultará em uma sensibilidade maior. Portanto, para uma dada consulta, o indivíduo que possui os maiores coeficientes de correlação entre os demais indivíduos que atendem à consulta, será o que causa o maior impacto no valor da sensibilidade. Além disso, quando os graus de correlação entre todos os pares de indivíduos se igualam a *zero*, evidencia-se um cenário onde não há correlação nos dados e, portanto, os resultados obtidos são iguais aos do modelo de Privacidade Diferencial tradicional.

Algumas estratégias foram sugeridas para identificar e mensurar a correlação entre os indivíduos, notadamente o Coeficiente de Correlação de *Pearson* que se encontra entre as mais promissoras, visto que se trata de um método robusto para identificar relacionamentos implícitos,

independentemente de qualquer conhecimento prévio sobre os dados. Entretanto, os autores deixam em aberto a utilização de novas estratégias que possam surgir para o processo de análise e descoberta de correlação.

Além dos graus de correlação, o trabalho impõe um *threshold* δ_0, o qual tem a finalidade de limitar a correlação existente entre os indivíduos, de forma que os indivíduos que possuem graus de correlação abaixo de δ_0 têm seus valores desconsiderados, ou seja, igualados a *zero*.

Por se tratar do trabalho que mais se aproxima do nosso, utilizando estratégias capazes de identificar relacionamentos implícitos entre os indivíduos sem a necessidade de qualquer conhecimento prévio sobre os dados, foi escolhido para fins de comparação com a nossa solução. Além disso, algumas definições apresentadas nesse trabalho foram reproduzidas em nossa solução.

3.1.4 Estratégia proposta em (LIU et al., 2016)

Este trabalho propõe um novo modelo de privacidade, denominado Privacidade Diferencial Dependente (*Dependent Differential Privacy*), que procura atacar o problema da privacidade em cenários onde os indivíduos de um conjunto de dados possuem relações de dependência entre si. O novo modelo de privacidade tem suas propriedades garantidas através de um novo mecanismo, também proposto pelos autores, denominado mecanismo de perturbação dependente (*Dependent Perturbation Mechanism*).

Nesse modelo de privacidade, dois novos parâmetros são acrescentados, os parâmetros $\hat{L}$ e $\hat{D}$. O parâmetro $\hat{L}$ representa o número máximo de dependências existente entre os indivíduos. Em outras palavras, qualquer indivíduo possui relações de dependência com, no máximo, outros $\hat{L}-1$ indivíduos. Por sua vez, o parâmetro $\hat{D}$ diz respeito às probabilidades de dependência existentes entre os $\hat{L}$ indivíduos dependentes entre si.

Os parâmetros $\hat{L}$ e $\hat{D}$ são empregados na definição de conjuntos de dados vizinhos, a qual será empregada em uma nova definição de sensibilidade para conjuntos de dados com relações de dependência, denominada sensibilidade dependente. Um conjunto de dados vizinhos deixa de ser aquele que apenas difere em um elemento do conjunto de dados original e passar a ser aquele que, além de diferir em um indivíduo, a ausência desse indivíduo causa mudanças nos valores de outros $\hat{L}-1$ indivíduos em função das probabilidades de dependência $\hat{D}$ presentes no relacionamentos entre os indivíduos.

Diante da nova definição de conjuntos de dados vizinhos, a sensibilidade dependente pode ser então calculada. Sua computação ocorre de maneira similar ao cálculo da sensibilidade tradicional, porém com o acréscimo do coeficiente de dependência, que multiplica o valor da sensibilidade obtida ao executar uma consulta sobre o conjunto de dados, considerando a nova definição de conjuntos de dados vizinhos. O coeficiente de dependência consiste no grau de relacionamento entre dois indivíduos, na perspectiva da privacidade, e seu valor varia entre 0 e 1. Para o valor do coeficiente igual a 1, identifica-se um cenário de dependência completa entre os indivíduos, de forma que um indivíduo pode ser unicamente determinado pelo outro.

Por fim, uma vez computada a sensibilidade dependente, o mecanismo de perturbação dependente utiliza o mecanismo de *Laplace*, através do ruído de *Laplace*, para gerar o ruído que será acrescido à resposta real da consulta.

3.2 Privacidade Diferencial e Agrupamento de Dados

Inicialmente proposta para ser um modelo interativo, a Privacidade Diferencial foi adaptada ao longo do tempo para atuar, também, como um modelo de privacidade sintático para a publicação de dados. Entretanto, por não ser, a princípio, um modelo sintático, os dados publicados podem apresentar excesso de ruído, comprometendo a utilidade dos dados. Dessa forma, alguns trabalhos (SORIA-COMAS *et al.*, 2014; SÁNCHEZ *et al.*, 2016), buscando melhorar a utilidade dos dados, combinaram as vantagens dos modelos interativos e sintáticos ao utilizar os modelos de Privacidade Diferencial e *k*-anonimato no processo de publicação de dados. O modelo *k*-anonimato, atuando como um algoritmo de agrupamento, reduz a sensibilidade aplicada sobre o mecanismo de Privacidade Diferencial e a quantidade de ruído adicionada sobre os dados publicados, aumentando a utilidade dos mesmos. No entanto, uma vez que técnicas de agrupamento são amplamente empregadas para identificar objetos similares entre si, acreditamos ser uma estratégia viável no processo de identificação de relacionamentos entre indivíduos, aumentando, dessa forma, a utilidade dos dados perante os trabalhos apresentados na Seção 3.1.

3.2.1 Estratégia proposta em (SORIA-COMAS et al., 2014)

Este trabalho apresenta uma estratégia para a publicação de dados que combina dois modelos de privacidade, o *k*-anonimato e a Privacidade Diferencial, com o objetivo de diminuir a quantidade de ruído acrescida aos dados no momento de sua publicação.

O princípio básico do trabalho consiste em aplicar uma técnica de micro-agregação para agrupar o conjunto de dados original, tornando-o um conjunto de dados k-anônimo, e um mecanismo diferencialmente privado sobre esse conjunto, levando em consideração os grupos formados e seus respectivos centroides. Dessa forma, a quantidade de ruído adicionada aos dados, baseando na disposição dos grupos e seus centroides, será consideravelmente menor do que quando se considera os dados em sua totalidade. Isso ocorre devido à sensibilidade dentro dos *clusters* ser menor do que a sensibilidade quando se considera o conjunto de dados como um todo.

A estratégia proposta divide-se, basicamente, em duas etapas: uma primeira que agrupa os dados através de uma técnica de micro-agregação e uma segunda que adiciona ruído aos indivíduos de cada *cluster* por meio do mecanismo de *Laplace*.

Na primeira etapa, os autores utilizam uma técnica de micro-agregação insensível, com o intuito de formar *clusters* menos sensíveis a mudanças. Em uma micro-agregação insensível, a mudança de um indivíduo de um *cluster* irá afetar, no máximo, um indivíduo em cada um dos outros *clusters*. Caso a estratégia de micro-agregação insensível não tivesse sido empregada, qualquer mínima alteração no conjunto de dados, ou seja, uma modificação em apenas um indivíduo, poderia ocasionar uma mudança abrupta na formação dos grupos. Nessa etapa, os registros são, inicialmente, ordenados, através de uma relação de ordem total, onde todos os atributos dos registros são considerados na ordenação. Uma vez ordenados, os *clusters* passam a ser populados, sequencialmente, de forma que todos os *clusters* possuem tamanho k, com exceção do último, que pode possuir um tamanho de até $2k-1$. Por fim, calcula-se o centroide de cada *cluster* e generaliza-se os valores de cada elemento presente no *cluster* pelo valor do centroide.

A Figura 9 exemplifica a etapa de micro-agregação sobre um conjunto de dados hipotético com parâmetro $k = 5$. O lado esquerdo representa o conjunto de dados original e o lado direito o mesmo conjunto após a modificação de apenas um elemento. Como pode ser observado, por não se tratar de uma estratégia insensível, os centroides (representados por um $\mathbf{x}$) dos *clusters* foram completamente modificados.

Em contrapartida, a Figura 10 apresenta a etapa de micro-agregação insensível, onde é possível observar que, para o mesmo conjunto de dados da Figura 9, os centroides dos *clusters* permanecem similares mesmo após a modificação de um elemento. A utilização de uma estratégia insensível contribui, também, para manter o máximo de homogeneidade dentro dos

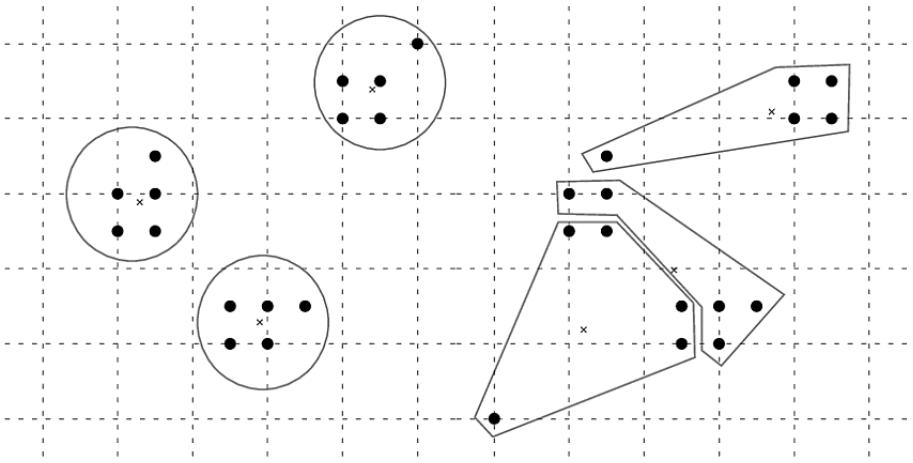


Figura 9 – *Clusters* e centroides gerados a partir do algoritmo de micro-agregação com $k = 5$. À esquerda o conjunto de dados original e à direita o mesmo conjunto de dados com um elemento modificado (DOMINGO-FERRER *et al.*, 2016).

clusters.

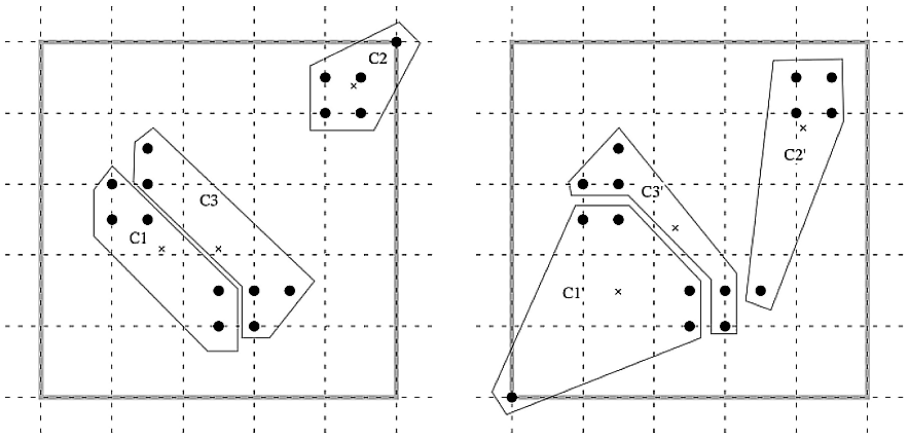


Figura 10 – *Clusters* e centroides gerados a partir do algoritmo de micro-agregação insensível com $k = 5$. À esquerda o conjunto de dados original e à direita o mesmo conjunto de dados com um elemento modificado (DOMINGO-FERRER *et al.*, 2016).

Na segunda etapa, os *clusters* provenientes da etapa de micro-agregação são utilizados para calcular a sensibilidade que será empregada na geração de ruído pelo mecanismo de *Laplace*. Por se tratar de uma abordagem de publicação, não havendo consultas sobre o conjunto de dados, a sensibilidade passa a ser dada em função da máxima variação entre os elementos no domínio do conjunto de dados, sendo denotada por $\Delta(X)$. No entanto, a modificação de um elemento em cada um dos *clusters* resultará em um mudança de, no máximo, $\frac{\Delta(X)}{k}$ nos valores dos centroides. Portanto, visto que, no pior cenário, cada centroide terá seu valor modificado em um fator de $\frac{\Delta(X)}{k}$, a sensibilidade resultante, a qual será utilizada no processo de geração do ruído de *Laplace*, é dada por $\frac{N}{k} \times \frac{\Delta(X)}{k}$, onde N representa o número de indivíduos do conjunto

de dados X e $\frac{N}{k}$ representa o número de *clusters* gerados pelo algoritmo de micro-agregação.

Por fim, em posse do valor da sensibilidade, uma amostra de ruído, proveniente do mecanismo de *Laplace*, é gerada e adicionada a cada elemento pertencente aos *clusters*, publicando-os posteriormente.

3.2.2 Estratégia proposta em (SÁNCHEZ et al., 2016)

Por fim, este trabalho consiste em um aperfeiçoamento do último trabalho dos autores (SORIA-COMAS *et al.*, 2014), com o objetivo de melhorar a utilidade dos dados publicados, através da diminuição da sensibilidade. Assim como no trabalho anterior, os autores continuam mantendo o contexto de publicação de dados utilizando a combinação de dois modelos de privacidade, o k-anonimato e a Privacidade Diferencial

Diferente do primeiro trabalho, o qual agrupa o conjunto de dados através de um algoritmo de micro-agregação a nível de registro, ou seja, onde todos os atributos dos indivíduos são considerados durante o agrupamento, nesse trabalho considera-se os atributos individualmente, de forma que a micro-agregação é realizada a nível de atributo. Assim, são realizadas m micro-agregações distintas, onde m é número de atributos existentes.

A micro-agregação multivariada, i.e., realizada a nível de registro, por mais que insensível, não garante a propriedade da ordem total, podendo ocasionar o caso extremo de se modificar um elemento de cada *cluster* e o valor de seus respectivos centroides, resultando em uma sensibilidade de $\frac{N}{k} \times \frac{\Delta(X)}{k}$. Por sua vez, uma micro-agregação univariada, i.e., realizada a nível de atributo, não apresenta a mesma limitação, visto que os elementos referentes a um mesmo atributo pertencem a um mesmo domínio e já se encontram ordenados, provendo a propriedade de ordem total. Dessa forma, a mudança de um elemento não mais resultará em modificações de todos os *clusters* e centroides.

As sensibilidades passam a ser dadas em função dos m atributos, de forma que cada atributo distinto possui sua própria sensibilidade. A sensibilidade $\Delta(A_i)$ representa a sensibilidade do atributo A_i, sendo $i \leq m$, a qual mensura a maior diferença observada entre os valores do atributo A_i, no domínio do conjunto de dados X.

Uma vez calculadas as sensibilidades, o processo de publicação se assemelha ao trabalho anterior, com o diferencial de que, agora, cada atributo publicado possui sua própria sensibilidade. Dessa forma, para cada elemento que compõe o registro de um determinado indivíduo, onde seu valor foi generalizado pelo valor do centroide do *cluster* ao qual pertence,

um ruído de *Laplace*, com sensibilidade $\frac{\Delta(A_i)}{k}$ e escala $\frac{\Delta(A_i)}{k \times \varepsilon}$, é adicionado ao seu valor.

3.3 Discussão

Os trabalhos apresentados neste capítulo dividem-se em abordagens interativas e não-interativas, ou ambas, dependendo do trabalho. As abordagens, em sua grande maioria, buscam garantir as propriedades do modelo de Privacidade Diferencial, com exceção dos dois trabalhos que utilizam técnicas de micro-agregação para a publicação dos dados, os quais empregam, também, o modelo k-anonimato. Mecanismos padrões, como o de *Laplace* e o Exponencial, foram os mais utilizados, salvo alguns trabalhos que implementaram seus próprios mecanismos de privacidade, sendo empregados sobre diferentes tipos de dados. Por fim, diferentes estratégias para mensurar a correlação, ou dependência, entre os indivíduos foram empregadas nos trabalhos direcionados a conjunto de dados onde evidencia-se o relacionamento entre indivíduos.

Dentre as limitações da estratégia proposta em (KIFER; MACHANAVAJJHALA, 2014), pode-se evidenciar a necessidade de conhecimento pleno sobre o domínio da aplicação, i.e., sobre os dados, de maneira a especificar os três conjuntos de requisitos do *framework*, o que torna a solução um pouco difícil de ser implementada e, de certa forma, impraticável. Os autores não deixam claro quanto ao ambiente de aplicação do *framework*, deixando a entender que o mesmo pode ser utilizando tanto em ambientes interativos quanto não-interativos. Além disso, nenhum mecanismo que possa ser aplicado é proposto.

A estratégia proposta em (CHEN *et al.*, 2014) apresenta uma limitação no que diz respeito à medida de correlação, a qual é mensurada em função do número máximo de conexões entre os indivíduos do grafo. Dessa forma, por mais que os autores tenham modificado o mecanismo Exponencial, com o intuito de reduzir a quantidade de ruído adicionada ao grafo anonimizado publicado, a utilidade foi drasticamente afetada. Por sua vez, a estratégia proposta em (LIU *et al.*, 2016) possui limitações semelhantes no que diz respeito à medida de correlação, mesmo que aplicada sobre tipos de dados distintos. Ainda assim, a correlação é dada em função do número de indivíduos dependentes no conjunto de dados e suas respectivas probabilidades. No entanto, não existe uma forma autônoma de estimar esses valores caso os dados tenham sido gerados de maneira desconhecida, como ocorre na maioria das vezes. Dessa forma, se torna necessário o conhecimento sobre os dados a fim de estimar tais probabilidades. Por se tratar de um método falho, o mecanismo proposto no trabalho necessita passar por um relaxamento, para cobrir possíveis falhas.

O trabalho proposto em (ZHU *et al.*, 2015) supre a necessidade do conhecimento prévio sobre os dados ao propor uma estratégia que independe de qualquer conhecimento para identificar e mensurar a correlação entre os indivíduos. Sua estratégia destaca o coeficiente de correlação de *Pearson* como alternativa para medir a correlação, embora os autores deixem em aberto a utilização de outra estratégia que venha a surgir. Além disso, mudanças no cálculo da sensibilidade foram introduzidas a fim de garantir a privacidade dos indivíduos em conjuntos de dados correlacionados, passando a levar em consideração o grau de correlação entre os indivíduos. No entanto, a estratégia acrescenta muito ruído às respostas publicadas, visto que os autores, ao calcular a correlação entre todos os pares de indivíduos, consideram que todos os indivíduos podem estar correlacionados. Além disso, a estratégia se torna demasiadamente custosa, visto que é necessário mensurar a correlação entre quaisquer pares de indivíduos. Por fim, os autores propõe um novo mecanismo de privacidade, mas também deixam em aberto a utilização de qualquer outro mecanismo.

Por sua vez, a estratégias apresentadas nos trabalhos (SORIA-COMAS *et al.*, 2014; SÁNCHEZ *et al.*, 2016), embora utilizem uma técnica de micro-agregação para agrupar o conjunto de dados, não buscam tratar o problema da correlação dos dados. O foco do trabalho é, exclusivamente, agrupar o conjunto de dados com o objetivo de reduzir a quantidade do ruído sobre os dados publicados. Além disso, os trabalhos possuem limitações quanto ao valor mínimo do parâmetro k, o qual especifica o tamanho mínimo dos *clusters*. O trabalho de (SORIA-COMAS *et al.*, 2014), por empregar uma técnica de micro-agregação a nível de registro, requer um parâmetro $k > \sqrt{N}$, sendo N o tamanho do conjunto de dados, enquanto que o trabalho de (SÁNCHEZ *et al.*, 2016), o qual emprega uma técnica de micro-agregação a nível de atributo, requer que o valor do parâmetro k seja, ao menos, igual a 2. No entanto, se tratando de um cenário onde há dados correlacionados, a limitação do parâmetro k torna as estratégias inadequadas, visto que cada indivíduo é forçado a estar correlacionado a, pelo menos, $k-1$ outros indivíduos. Além disso, nenhuma das duas técnicas é capaz de identificar *outliers*, i.e., indivíduos de que não se correlacionam com outros indivíduos, uma vez que não faz sentido ter um valor de $k = 1$.

A nossa contribuição adota uma abordagem interativa do modelo de Privacidade Diferencial, garantindo suas propriedades através do mecanismo de *Laplace*, sobre conjunto de dados correlacionados em formato tabular. A partir desse modelo, propomos um método que previne ataques probabilísticos, por parte de um usuário malicioso, através da publicação de respostas anonimizadas.

Diferente dos trabalhos apresentados, nossa solução emprega técnicas de agrupamento a fim de identificar apenas os indivíduos potencialmente correlacionados, em oposição àqueles que consideram que todos os indivíduos se relacionam entre si, e, assim, reduzir a quantidade de ruído acrescida às respostas e aumentar a utilidade dos dados providos. Além disso, utilizamos uma técnica de seleção de atributos para manter apenas os atributos mais relevantes para o processo de identificação de relacionamentos, o que garante um resultado mais preciso e com melhor desempenho. Por fim, propomos uma estratégia probabilística e autônoma para mensurar a correlação entre os indivíduos, a qual é independente de qualquer conhecimento prévio sobre os dados.

A Tabela 8 traz um resumo comparativo entre as abordagens propostas e a nossa contribuição (MENDONÇA *et al.*, 2017).

Tabela 8 – Análise comparativa entre os trabalhos relacionados.

Trabalho	Abordagem	Tipo de Dado	Modelo de Priv.	Mecanismo de Priv.	Medida de Correlação
Kifer; Machanavajjhala, 2014	Interativa Não-interativa	Relacional	PD	Customizável	Regras determinadas pelo analista
Chen et al., 2014	Não-interativa	Grafos	PD	Exponencial	Número de conexões entre os indivíduos
Zhu et al., 2015	Interativa	Relacional	PD	*Laplace**	Coeficiente de correlação de *Pearson**
Liu et al., 2016	Interativa Não-interativa	Relacional	PD	Perturbação Dependente	Numero de indivíduos dependentes Dependência probabilística entre os indivíduos
Soria-Comas et al., 2014	Não-interativa	Relacional	*k*-anonimato + PD	*Laplace*	-
Sánchez et al., 2016	Não-interativa	Relacional	*k*-anonimato + PD	*Laplace*	-
Mendonça et al., 2017	Interativa	Relacional	PD	*Laplace*	Coeficiente de correlação de *Spearman* Relacionamentos dados por *clusters*

3.4 Conclusão

Este capítulo apresentou os principais trabalhos relacionados com o tema desta dissertação. Embora existam diversas abordagens que visam proteger a privacidade dos indivíduos em dados correlacionados, tais trabalhos não propõem estratégias viáveis, tornando o processo de identificação dos relacionamentos entre indivíduos muito complexo ou adicionando muito ruído às soluções, comprometendo drasticamente a utilidade dos dados. A contribuição apresentada nesta dissertação é comparada com o trabalho relacionado (ZHU *et al.*, 2015), uma vez que ambas as abordagens empregam o modelo de Privacidade Diferencial sobre dados correlacionados e identificam os relacionamentos implícitos existentes no conjunto de dados. Além disso, os respectivos graus de correlação são mensurados sem a necessidade de qualquer conhecimento prévio sobre os dados, visando impedir ataques probabilísticos, executados por usuários maliciosos, contra os indivíduos. A descrição completa da nossa abordagem e seus resultados experimentais serão detalhados no capítulo seguinte desta dissertação.

4 MÉTODO PROPOSTO

4.1 Visão Geral

Este capítulo apresenta o conjunto de soluções propostas, e seus respectivos resultados, com o objetivo de solucionar o problema de preservação de privacidade em conjuntos de dados correlacionados em ambientes interativos. Para isso, propomos três abordagens que utilizam técnicas de agrupamento com a finalidade de identificar os indivíduos correlacionados. De posse dessas informações de correlação, aplica-se o modelo de Privacidade Diferencial com maior garantia contra ataques probabilísticos dos indivíduos envolvidos. Isto é necessário em oposição à aplicação teórica que não leva em conta a correlação entre os dados dos indivíduos. É importante salientar que a utilização de técnicas de agrupamento durante o processo de descoberta de relacionamento entre indivíduos é crucial para manter a utilidade dos dados gerados. Além disso, propomos uma solução com um excelente ganho de desempenho perante os demais trabalhos existentes na literatura.

Em nossas abordagens, utilizamos algumas definições propostas no trabalho que define a Privacidade Diferencial Correlacionada (ZHU *et al.*, 2015), como: Matriz de Graus de Correlação (Δ) e Sensibilidade Correlacionada (CS_f). As três abordagens propostas são compostas por duas grandes etapas, sendo elas:

1. **Construção da Matriz de Graus de Correlação:** agrupa o conjunto de dados D utilizando algum algoritmo de agrupamento e constrói a matriz Δ baseando-se na disposição dos *clusters* e em alguma medida de correlação;
2. **Mecanismo de *Laplace* Correlacionado:** calcula a sensibilidade correlacionada utilizando a matriz Δ recém construída e retorna respostas diferencialmente privadas utilizando o mecanismo de *Laplace*.

A segunda etapa é comum às três abordagens, enquanto a execução da primeira etapa varia de acordo com a abordagem empregada. Para a primeira etapa temos as seguintes variações:

- Construção da Matriz de Graus de Correlação com *DBSCAN* **(Abordagem 1)**: utiliza como técnica de agrupamento o algoritmo *DBSCAN*;
- Construção da Matriz de Graus de Correlação com *GMM* **(Abordagem 2)**: utiliza o algoritmo de agrupamento probabilístico *GMM*;
- Construção da Matriz de Graus de Correlação com *GMM* e Redução de Dimensi-

onalidade **(Abordagem 3)**: além de focar na utilidade dos dados, enfatiza-se o desempenho empregando uma técnica de redução de dimensionalidade antes de agrupar os dados através do algoritmo *GMM*, visto que o processo de identificar os indivíduos correlacionados e mensurar seus respectivos graus de correlação é demasiadamente custoso.

O Algoritmo 1 apresenta a estrutura base do algoritmo utilizado nas três abordagens. O algoritmo tem como entrada um conjunto de dados D, uma consulta f, o orçamento de privacidade ε e um algoritmo de agrupamento, denominado $algo\_agrup$, juntamente com seus parâmetros de configuração, definidos por *params*. A saída do algoritmo é uma resposta anonimizada $f'(D)$ da consulta f aplicada sobre o conjunto de dados correlacionado D que obedece ao modelo de Privacidade Diferencial, garantindo que um usuário malicioso não será bem sucedido em um ataque probabilístico.

Primeiramente, o Algoritmo 1 executa a etapa 1 (linha 2), onde a matriz Δ é construída. Nessa etapa, o algoritmo agrupa o conjunto de dados D utilizando algum algoritmo de clusterização $algo\_agrup[params]$ para formação de *clusters*. Os *clusters* determinam os indivíduos que estão correlacionados entre si e, a partir disso, é possível mensurar seus respectivos graus de correlação e construir a matriz Δ. Como dito anteriormente nesta seção, cada abordagem implementa essa etapa de acordo com suas especificidades (detalhadas na Seção 4.3).

Na etapa 2 (linha 3), utiliza-se a recém construída matriz Δ, o conjunto de dados D, a consulta f e o parâmetro ε para executar o mecanismo de *Laplace* e obter a resposta anonimizada $f'(D)$. O Algoritmo 2 representa a etapa 2, visto que se trata de uma etapa comum às três abordagens, ou seja, que não sofre alterações. Nessa etapa, calcula-se a *sensibilidade* correlacionada (linha 2) para gerar a quantidade de ruído (linha 3) que será adicionada à resposta real da consulta (linha 4) antes de ser retornada pelo algoritmo (linha 5). Por fim, o Algoritmo 1 finaliza sua execução ao retornar a mesma resposta anonimizada $f'(D)$ (linha 4) obtida na etapa 2.

4.2 Configuração Experimental

Antes de partimos para a explicação das abordagens e seus respectivos resultados, apresentaremos o ambiente de desenvolvimento, os conjuntos de dados e as métricas utilizadas para avaliar a utilidade e desempenho das três diferentes abordagens e do trabalho de Zhu et al., que será chamado, por praticidade, de *baseline*.

Algoritmo 1: Método Base

Entrada: Conjunto de dados D, consulta f, orçamento de privacidade ε, algoritmo de agrupamento $algo\_agrup$ juntamente com seus parâmetros $params$ de configuração

Saída: Resposta anonimizada $f'(D)$ da consulta f aplicada sobre o conjunto de dados D

início
  // Etapa 1: Agrupamento dos dados e construção da matriz de graus de correlação Δ
  $\Delta \leftarrow \text{ETAPA\_1}(D, algo\_agrup[params])$;
  // Etapa 2: Execução do Mecanismo de *Laplace* Correlacionado
  $f'(D) \leftarrow \text{ETAPA\_2}(D, f, \Delta, \varepsilon)$;
  retorna $f'(D)$;
fim

Algoritmo 2: ETAPA_2

Entrada: Conjunto de dados D, consulta f, orçamento de privacidade ε, matriz de graus de correlação Δ

Saída: Resposta anonimizada $f'(D)$ da consulta f aplicada sobre o conjunto de dados D

início
  $sensibilidade \leftarrow CS_f(D, f, \Delta)$;
  $ruido \leftarrow Laplace(0, \frac{sensibilidade}{\varepsilon})$;
  $f'(D) \leftarrow f(D) + ruido$;
  retorna $f'(D)$;
fim

4.2.1 Ambiente de Desenvolvimento

As abordagens propostas e o *baseline* foram implementados na linguagem Python 2.7, utilizando os pacotes Scipy 0.17 (JONES *et al.*, 2001), scikit-lean 0.18.1 (PEDREGOSA *et al.*, 2011) e pandas 0.13.1 (MCKINNEY, 2010). Todos os experimentos foram executados em uma máquina *desktop*, pertencente ao Laboratório de Sistemas e Banco de Dados (LSBD/UFC), equipada com sistema operacional Ubuntu 14.04, processador Intel Core i5 de 3,2 GHz, 8 GB de RAM e disco de 500 GB.

4.2.2 Conjuntos de Dados

Foram utilizados três conjuntos de dados reais do *UCI Machine Learning Repository* (LICHMAN, 2013). Os conjuntos consistem em dados de censo, informações de pagamentos e comportamento humano, e estão, respectivamente, detalhados na Tabela 9.

Tabela 9 – Conjuntos de dados utilizados nos experimentos.

Conjunto de Dados	Número de Registros	Número de Atributos
Adult	30.162[1]	14
CCC[2]	30.000	24
SBRHAPT[3]	10.929	561

4.2.3 Análise de Utilidade

Propomos duas medidas distintas com o objetivo de avaliar a utilidade dos dados provenientes de nossas abordagens e do *baseline*, denominadas Análise do Erro Relativo e *Trade-off* Utilidade-Privacidade. Nelas, foram empregadas 100 consultas do tipo **COUNT**, embora qualquer outro tipo de consulta de agregação pudesse ter sido utilizada sem qualquer prejuízo, todas geradas aleatoriamente, com o objetivo de tornar os resultados mais acurados. Além disso, garantimos as propriedades da Privacidade Diferencial através do mecanismo de *Laplace*.

4.2.3.1 Análise do Erro Relativo

Nessa análise, empregamos uma variante do Erro Relativo (SHOARAN *et al.*, 2012; XIAO *et al.*, 2011; CHEN *et al.*, 2014) como métrica para avaliar a utilidade dos dados, a qual denominamos Erro Relativo Médio (ERM) (XIAO *et al.*, 2011). Através do ERM é possível mensurar a discrepância entre as respostas reais das consultas e suas versões anonimizadas, que são providas por um mecanismo diferencialmente privado.

Definição 10 *Dado uma consulta f, seja $f(D)$ a resposta real da consulta f aplicada sobre o conjunto de dados D e $f'(D)$ a resposta real da consulta acrescida de um ruído gerado a partir de um mecanismo diferencialmente privado, o Erro Relativo (ER) é definido por:*

$$ER = \frac{|f'(D) - f(D)|}{f(D)}.$$

Uma vez definido o erro relativo, o erro relativo médio é dado por:

$$ERM = \frac{\sum_{i=1}^{n} \frac{|f_i'(D) - f_i(D)|}{f_i(D)}}{n},$$

[1] Composto, inicialmente, por 48.842 registros. No entanto, restaram apenas 30.162 após a remoção dos registros com atributos vazios.

[2] Abreviação para *Credit Card Clients*.

[3] Abreviação para *Smartphone-Based Recognition of Human Activities and Postural Transitions*.

onde n é o número de consultas geradas e $f_i(D)$ e $f_i'(D)$ são, respectivamente, a resposta real e sua versão anonimizada, por meio do mecanismo de *Laplace*, da i-ésima consulta.

O ERM se torna preferível em relação ao ER visto que, devido a aleatoriedade do mecanismo de *Laplace*, é possível, sob o mesmo ε, que uma consulta com uma sensibilidade maior que a de outra consulta retorne um ruído menor, por mais que a probabilidade seja baixa. Portanto, o ERM se torna uma estratégia para minimizar esse tipo de ocorrência.

Utilizamos o ERM para avaliar o impacto dos parâmetros de configuração dos algoritmos de agrupamento e do parâmetro ε do modelo de Privacidade Diferencial sobre as saídas do mecanismo. É necessário fixar o valor do parâmetro ε para avaliar o impacto dos parâmetros de configuração dos algoritmos de agrupamento, e vice-versa.

4.2.3.2 Trade-off Utilidade-Privacidade

Diferente da Análise do Erro Relativo, onde se avalia a diferença entre as respostas reais e suas versões anonimizadas em níveis percentuais, nosso interesse nessa análise consiste apenas em apresentar graficamente o quão distante as respostas reais estão de suas versões anonimizadas.

4.2.4 Análise de Desempenho

Essa análise consiste em mensurar o tempo decorrido (em minutos) para a construção da matriz de graus de correlação Δ. Nesse processo, consideram-se os tempos empenhados durante as etapas de identificar os indivíduos que se relacionam entre si e mensurar seus respectivos graus de correlação.

O tempo decorrido durante a etapa de execução do mecanismo de *Laplace* foi desconsiderado devido à sua imprevisibilidade, ou seja, o tempo necessário para calcular a sensibilidade de uma consulta varia em função da natureza da própria consulta e não dos dados propriamente ditos, podendo vir a distorcer os resultados.

Por se tratar de um ambiente interativo, o qual usuários requisitam consultas a um conjunto de dados e estes, por meio de um mecanismo, atendem às requisições dos usuários, é desejável que o processo aconteça em tempo hábil. Portanto, é imprescindível a utilização de um algoritmo de anonimização que viabilize a utilização do ambiente de maneira interativa.

4.3 Abordagens

Nessa seção, apresentaremos, em detalhes, cada abordagem no que diz respeito às suas implementações referente à etapa 1 do Algoritmo 1. Uma vez apresentada, cada abordagem será avaliada experimentalmente ao final de cada seção. Por se tratar de uma linha sucessória, cada abordagem será comparada, apenas, à abordagem anterior, com exceção da avaliação de desempenho, onde todas as abordagens serão comparadas entre si, incluindo o *baseline*. Dessa forma, para as demais avaliações experimentais, a abordagem 1 será comparada ao *baseline*, a abordagem 2 à abordagem 1 e, por fim, a abordagem 3 à abordagem 2.

Como visto na na Seção 2.2.2.2, a solução proposta pelo *baseline* consiste em adaptar a maneira como a sensibilidade é calculada a fim de garantir a privacidade dos indivíduos em conjuntos de dados correlacionados através de um mecanismo diferencialmente privado. Para isso, é preciso, inicialmente, identificar onde há evidências de correlação entre indivíduos e mensurá-las. Uma possível solução consiste em utilizar o Coeficiente de Correlação de Postos de *Spearman*, o qual recebe dois indivíduos de entrada e retorna um valor, no intervalo de $[-1,1]$, correspondente à correlação entre eles.

Em um primeiro momento, a autora supõe que todos os indivíduos são correlacionados entre si e calcula suas respectivas correlações através do coeficiente de *Spearman*. No entanto, ela define um *threshold* δ_0, o qual é utilizado pra filtrar e desconsiderar todos os relacionamentos existentes abaixo desse valor, tornando a solução não muito viável, visto que algumas informações cruciais acerca dos relacionamentos estão sendo descartadas. As correlações influenciam diretamente a sensibilidade e, consequentemente, a quantidade de ruído adicionada à resposta real de uma consulta. Em outras palavras, quanto maior os graus de correlação entre os indivíduos, maior será a sensibilidade e a quantidade de ruído necessária para garantir a privacidade dos mesmos.

Diante da limitação expressa por parte do *threshold* δ_0, nossas abordagens consistem em identificar apenas os indivíduos potencialmente correlacionados por meio de técnicas de agrupamento, sem a necessidade de, primeiro, mensurar as correlações e depois, se necessário, descartá-las.

4.3.1 Abordagem 1: Construção da Matriz de Graus de Correlação com DBSCAN

Em nossa primeira solução para o problema, utilizamos o algoritmo de agrupamento *DBSCAN* para identificar os indivíduos correlacionados em um conjunto de dados. Os *clusters* provenientes do agrupamento representam os indivíduos que se correlacionam entre si. Neste caso, considera-se que apenas os indivíduos pertencentes a um mesmo grupo correlacionam-se entre si, ou seja, não há a possibilidade de relacionamento entre grupos distintos, por mais que estes venham a ter elementos similares.

Dentre as vantagens encontradas no algoritmo *DBSCAN* estão a capacidade de encontrar *clusters* de formatos arbitrários, a fácil configuração por meio de apenas dois parâmetros, eps e $minPoints$, e sua robustez contra $outliers$. Em nosso contexto de dados correlacionados, elementos classificados como *outliers* são aqueles que não pertencem a nenhum grupo e, portanto, não se correlacionam com nenhum outro elemento, a não ser consigo mesmo. Além disso, por se tratar de um contexto de dados correlacionados, estamos em busca de identificar o máximo de relacionamentos existentes. Para isso, o parâmetro $minPoints$ foi fixado em 2, visto que, dessa forma, podemos identificar grupos com, até mesmo, apenas 2 indivíduos.

O Algoritmo 3 apresenta a etapa 1 com o algoritmo *DBSCAN*. Os parâmetros de entrada são o conjunto de dados D e os parâmetros de configuração do algoritmo de agrupamento *DBSCAN*, eps e $minPoints$. Uma vez que os *clusters* foram formados a partir do algoritmo *DBSCAN* (linha 2), a matriz Δ começa a ser construída no *loop* destacado na linha 3. Nesse *loop*, cada *cluster* é percorrido e, para cada par de indivíduos presentes no *cluster*, seus respectivos graus de correlação são medidos através do coeficiente de *Spearman*. Observe que, dessa forma, apenas a correlação entre indivíduos de um mesmo grupo é medida, diferente da abordagem proposta pelo *baseline*, onde todas as combinações possíveis entre dois indivíduos são mensuradas.

Visto o funcionamento da abordagem 1, apresentaremos alguns experimentos a fim de validá-la. A Figura 11 e Figura 12 avaliam a utilidade dos métodos em função do Erro Relativo Médio.

Na Figura 11 avaliamos o impacto dos parâmetros de configuração do algoritmo *DBSCAN*. Como demonstrado nos gráficos, quanto maior o valor do parâmetro eps, maior o erro relativo. Isso ocorre devido à formação dos *clusters*, de forma que quanto menor o valor do eps, mais esparsos estes serão, e quanto maior, mais densos. Quanto mais densos forem os *clusters*, maior será a quantidade de indivíduos pertencentes a um mesmo *cluster*, de forma que a matriz Δ, responsável por armazenar os graus de correlação entre os indivíduos, resultante da abordagem

Algoritmo 3: ETAPA_1 (Abordagem 1)

Entrada: Conjunto de dados D, parâmetros de configuração eps e $minPoints$ do algoritmo *DBSCAN*

Saída: Matriz de graus de correlação Δ

```
1  início
2      clusters ← DBSCAN(D, eps, minPoints);
3      para cada cluster ∈ clusters faça
           // Indivíduo i
4          para cada i ∈ cluster faça
               // Indivíduo j
5              para cada j ∈ cluster faça
6                  δ_ij ← SPEARMAN(i, j);
7                  Δ_ij ← δ_ij;
8              fim
9          fim
10     fim
11     retorna Δ;
12 fim
```

1, será bastante semelhante, ou igual, à matriz Δ resultante do *baseline*. Vale ressaltar que, nesse experimento, o outro parâmetro de configuração do *DBSCAN*, *minPoints*, foi fixado em 2. O comportamento uniforme do *baseline* se deve a este não variar em função dos parâmetros de configuração do *DBSCAN*, visto que, no *baseline*, considera-se a existência de um único grande *cluster* com a presença de todos os indivíduos. Por fim, o parâmetro ε do modelo de Privacidade Diferencial, utilizado pelo mecanismo de *Laplace*, foi fixado em 1.0, por se tratar de um valor que retém boa utilidade e privacidade, simultaneamente.

Por sua vez, a Figura 12 avalia o impacto do parâmetro ε do modelo de Privacidade Diferencial. No entanto, para esse experimento, foi necessário fixarmos, também, um valor para o parâmetro eps do algoritmo *DBSCAN*, visto que estamos interessado em avaliar o parâmetro ε, o que pode não ser uma tarefa trivial. Determinamos os valores do parâmetro eps mediante consulta da Figura 11, onde escolhemos valores intermediários, de forma a gerar *clusters* nem muitos esparsos e nem muito densos. Os valores escolhidos foram 0.4, 0.5 e 5.0 para os conjuntos de dados *Adult*, *CCC* e *SBRHAPT*, respectivamente, e estão destacados em **negrito** na Figura 11. De posse dos valores de configuração do *DBSCAN*, passamos a analisar o parâmetro ε. Como é possível observar nos gráficos, devido à própria natureza do parâmetro, temos que quanto menor o seu valor, menor será a utilidade dos dados gerados e, consequentemente, maior o erro relativo médio. Sabendo que utilidade e privacidade são grandezas inversamente proporcionais,

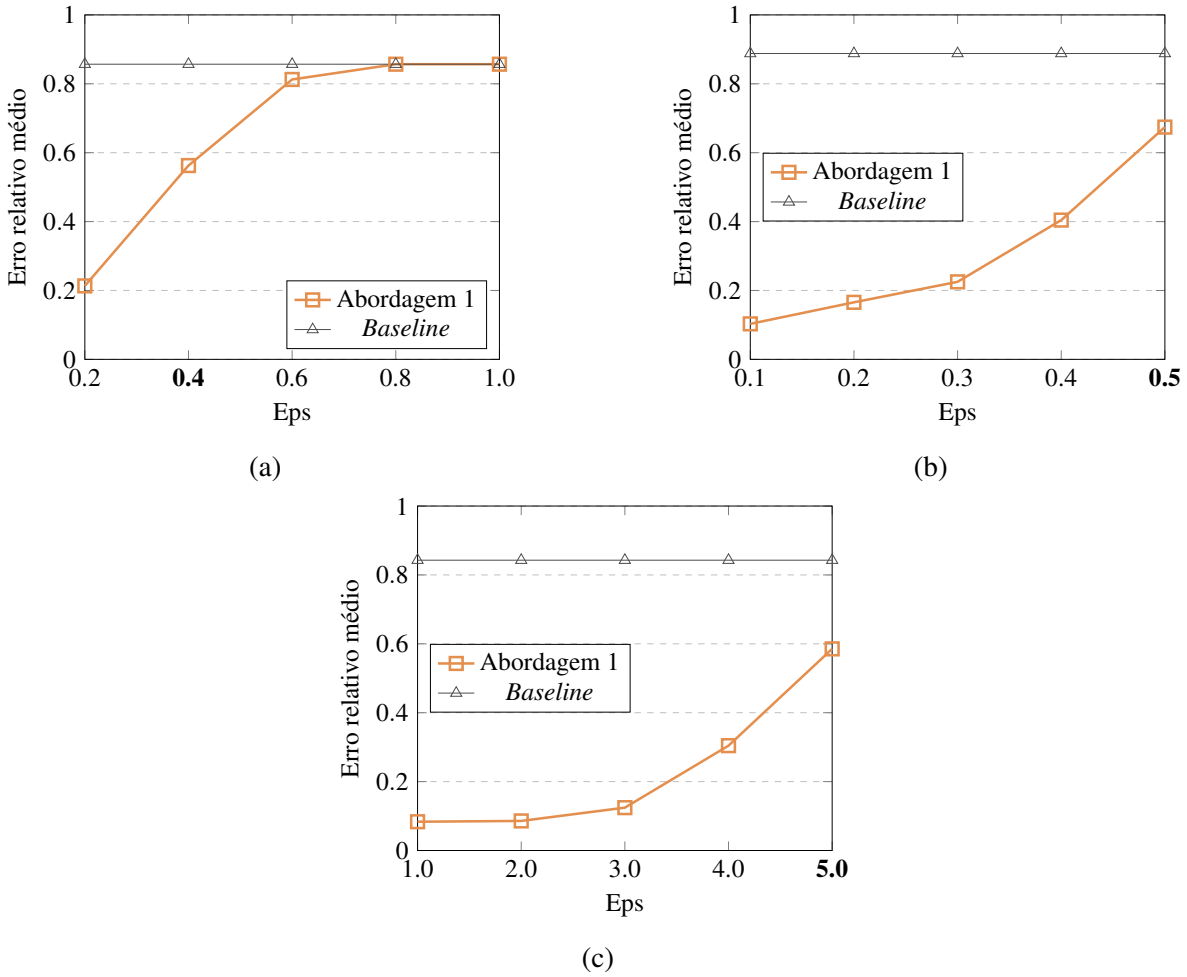


Figura 11 – Erro relativo resultante dos métodos destacados ao variar o parâmetro *eps* do algoritmo *DBSCAN*. (a) *Adult*. (b) *CCC*. (c) *SBRHAPT*.

valores menores de ε também garantem uma maior privacidade. Analogamente, temos que quanto maior o valor de ε, maior maior será a utilidade dos dados, menor o erro relativo médio e, consequentemente, menor a privacidade.

Por fim, na Figura 13, avaliamos o *trade-off* entre utilidade e privacidade ao comparar as respostas reais com suas versões anonimizadas, provenientes do mecanismo de *Laplace*. Esse gráfico proporciona uma melhor percepção das saídas do mecanismo, uma vez que o erro relativo médio proporciona um comparativo em termos percentuais, e não deixa tão claro o quão distante as respostas anonimizadas estão das respostas reais. Para esse experimento, repetimos os parâmetros de configuração utilizados no experimento anterior, representado pela Figura 12. Nos gráficos, as linhas diagonais representam as respostas reais das consultas. Portanto, quanto mais próximos da diagonal os resultados estiverem, mais úteis serão as informações providas pelo método. Como é possível observar, os resultados provenientes de nossa abordagem são substancialmente melhores que os resultados providos pelo *baseline*, visto que estão mais

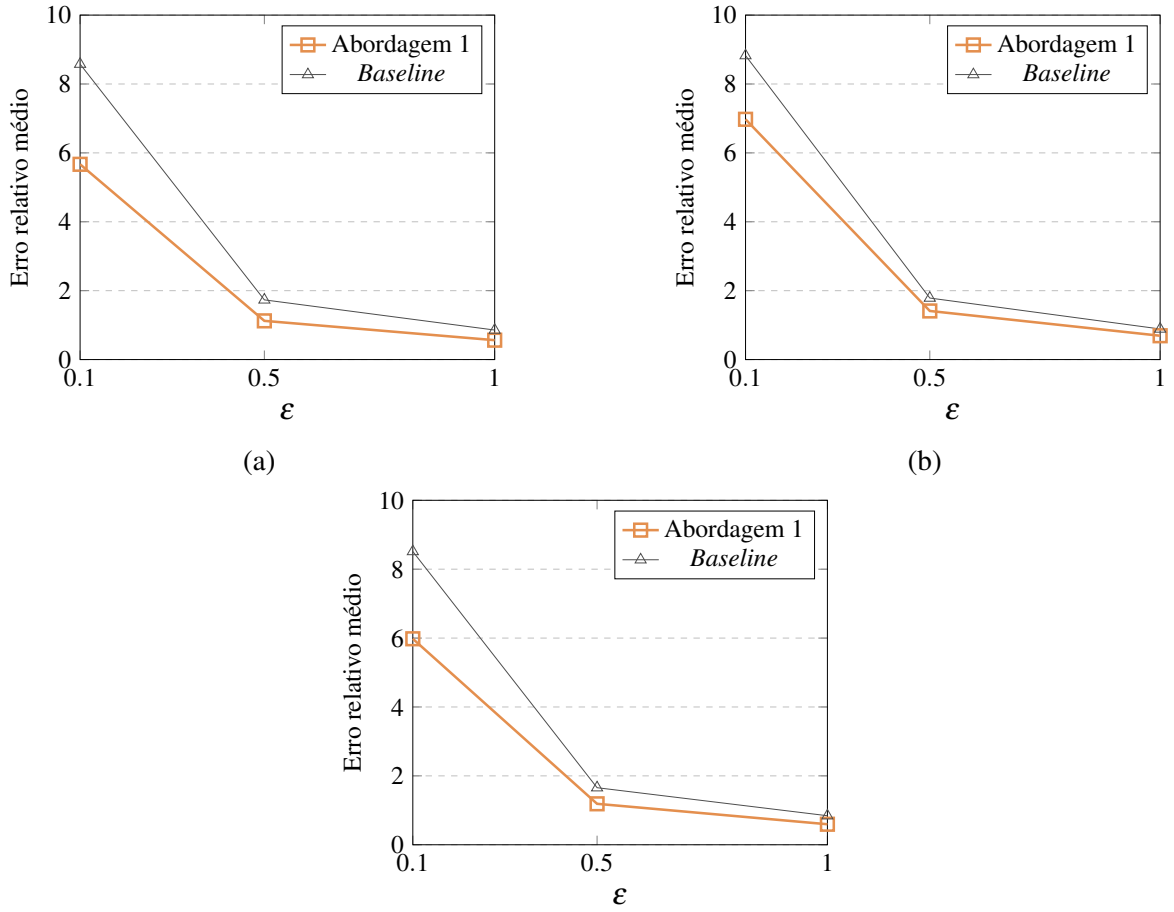


Figura 12 – Erro relativo resultante ao variar o parâmetro ε do modelo de Privacidade Diferencial. (a) *Adult*. (b) *CCC*. (c) *SBRHAPT*.

próximos da diagonal, assegurando, dessa forma, uma melhor utilidade do nosso método. A fim de facilitar o entendimento dos gráficos, consideramos apenas ruídos absolutos gerados pelo mecanismo de *Laplace*. Dessa forma, todas as respostas anonimizadas foram posicionadas acima da diagonal.

4.3.2 Abordagem 2: Construção da Matriz de Graus de Correlação com GMM

A abordagem 2 surgiu com o objetivo de solucionar algumas limitações apresentadas pela abordagem 1. Inicialmente, podemos citar a falta de flexibilidade apresentada pelo algoritmo *DBSCAN*, o qual não permite que um determinado indivíduo pertença a mais de um grupo, simultaneamente, por mais que o indivíduo possua características similares aos indivíduos de outros grupos. Além disso, embora o algoritmo *DBSCAN* exija poucos parâmetros em sua configuração, determiná-los não é uma tarefa trivial e acaba demandando um certo conhecimento prévio sobre os dados. Portanto, uma solução imediata consiste em determiná-los empiricamente.

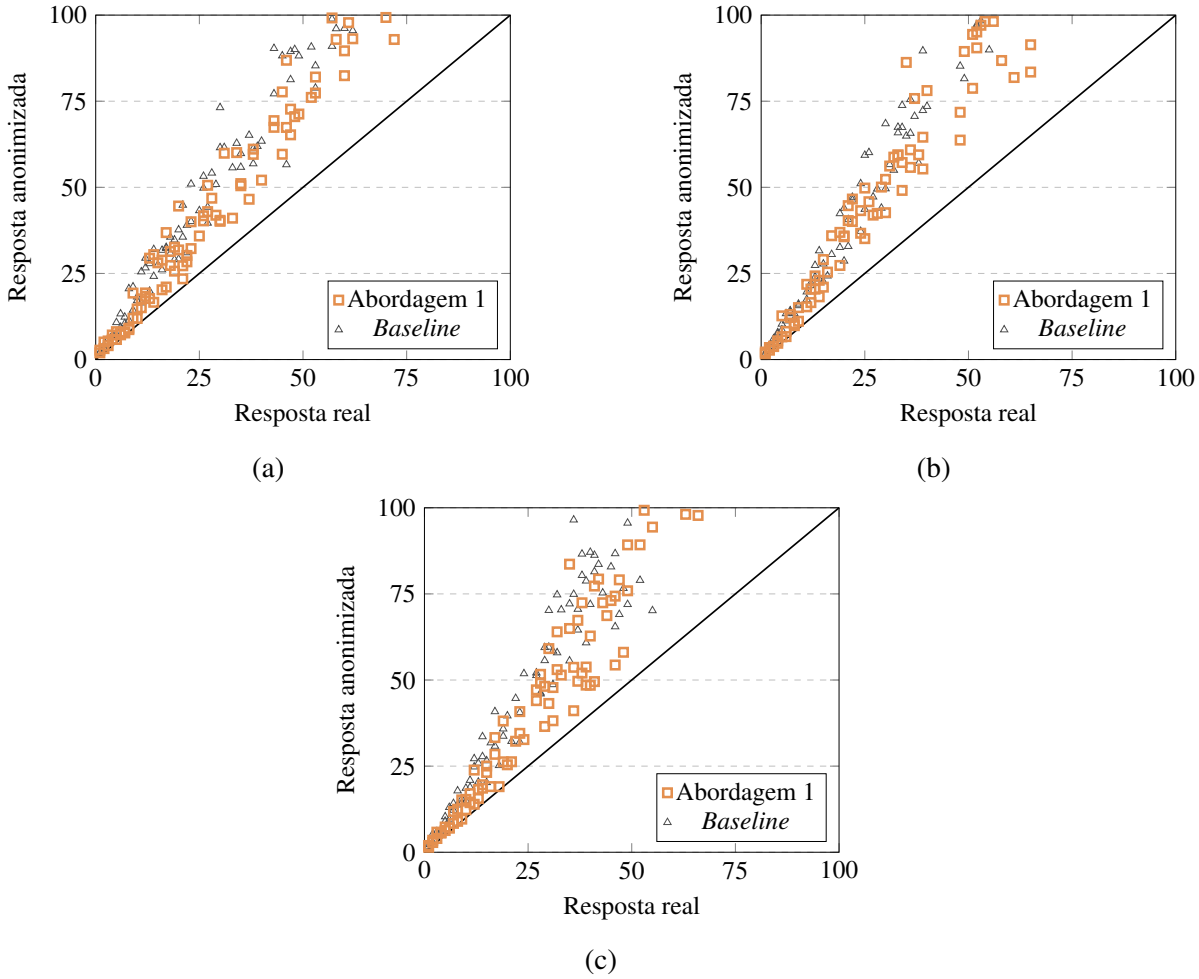


Figura 13 – Comparativo entre as respostas reais e suas versões anonimizadas. (a) *Adult*. (b) *CCC*. (c) *SBRHAPT*.

Para suprir as limitações recém mencionadas, o método empregado em nossa abordagem 2 consiste em utilizar o algoritmo *GMM*, em oposição ao algoritmo *DBSCAN*, para a tarefa de agrupar o conjunto de dados e identificar os relacionamentos existentes entre os indivíduos. O algoritmo *GMM* requer um único parâmetro de configuração, o número de *clusters* existentes. Isso poderia se tornar um grande problema e ser visto como uma limitação, visto que, na maioria dos casos, é preciso ter um conhecimento prévio sobre o domínio e a distribuição dos dados que estão sendo agrupados para determinar esse tipo de parâmetro e realizar um agrupamento coerente. Entretanto, diferente do algoritmo *DBSCAN*, o *GMM* possui artifícios que nos permitem estimar o número ótimo de *clusters* que representa o conjunto de dados, tornando desnecessário qualquer conhecimento prévio sobre os dados. O artifício empregado, a fim de descobrir o número ideal de *clusters*, é denominado *BIC*. O *BIC* é um critério que avalia a qualidade do agrupamento por meio de uma penalização que varia de acordo com os *clusters* gerados. Dessa forma, o agrupamento resultante com o menor *BIC* é aquele que melhor

representa o conjunto de dados.

Outra vantagem do algoritmo *GMM* consiste na formação dos *clusters* no que diz respeito à identificação dos relacionamentos existentes entre indivíduos. Ao contrário do algoritmo *DBSCAN*, que forma *clusters* nos quais os seus elementos pertencem exclusivamente àquele *cluster*, o algoritmo *GMM* parte de um princípio de incerteza. Neste, o conjunto de dados é agrupado a partir de uma mistura de várias gaussianas, de tal forma que um determinado indivíduo passa a possuir diferentes probabilidades de pertencer a cada um dos diferentes *clusters*, e não mais possui a certeza de pertencer a um único *cluster* exclusivamente.

De posse das probabilidades fornecidas pelo *GMM*, podemos calcular a probabilidade de quaisquer dois indivíduos estarem correlacionados, o que ocorre quando ambos pertencem ao mesmo *cluster*, independente de qual ele seja. Dessa forma, visto que, agora, a presença dos indivíduos em cada *cluster* é incerta, o cálculo do grau de correlação entre indivíduos deverá levar em consideração as suas respectivas probabilidades de pertencerem ao mesmo *cluster*. Entretanto, antes de modificarmos o cálculo do grau de correlação, é preciso definir a Matriz de Probabilidades Ω, a qual mantém as probabilidades dos indivíduos pertencerem aos mesmos *clusters*, sendo dada por:

Definição 11 *Dado um conjunto de dados D, um modelo $\mathscr{M}$ gerado a partir do algoritmo GMM e o número de clusters n, a Matriz de Probabilidades Ω, a qual mantém as probabilidades de dois indivíduos pertencerem a um mesmo cluster, é dada por:*

$$\Omega = \begin{bmatrix} \omega_{11} & \omega_{12} & \omega_{13} & \dots & \omega_{1n} \\ \omega_{21} & \omega_{22} & \omega_{23} & \dots & \omega_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \omega_{n1} & \omega_{n2} & \omega_{n3} & \dots & \omega_{nn} \end{bmatrix}$$

Onde $\mathscr{M}[i,k]$ representa a probabilidade do indivíduo i pertencer ao cluster k, onde $i \in D$ e $k \in [1,n]$. ω_{ij} é a probabilidade de dois indivíduos i e j pertencerem ao mesmo cluster, independente de qual cluster seja. A probabilidade ω_{ij} é dada por:

$$\omega_{ij} = \begin{cases} \sum_{k=1}^{n} \mathscr{M}[i,k] \times \mathscr{M}[j,k] & \text{se } i \neq j \\ 1.0 & \text{caso contrário} \end{cases}$$

A Lei da Probabilidade Total (PFEIFFER, 1978) e a própria Definição 9 asseguram a corretude da matriz Ω ao garantir que os valores assumidos por Ω jamais serão maiores que 1.0. Em outras palavras, a probabilidade de dois indivíduos pertencerem a um mesmo *cluster*, independente de qual ele seja, nunca excederá 100%.

Uma vez que o conjunto de dados tenha sido agrupado a partir do algoritmo *GMM* e as probabilidades dos indivíduos se relacionarem entre si foram identificadas, sendo representadas pela matriz Ω, os graus de correlação entre dois indivíduos podem ser computados a partir da multiplicação de suas respectivas probabilidades de esterem correlacionados pelo seus coeficientes de correlação de *Spearman*.

O Algoritmo 4 apresenta a etapa 1 com o algoritmo *GMM*. O algoritmo possui apenas o conjunto de dados *D* como um parâmetro de entrada. Inicialmente, o número de *clusters* a ser construído pelo modelo *GMM* que melhor agrupa e representa o conjunto *D* é identificado através do artifício *BIC* (linha 2). Nessa instrução são gerados 100 modelos distintos a partir do algoritmo *GMM*, variando o número de *clusters* de cada modelo de 1 a 100 e retornando o número de *clusters* responsável por construir o modelo com o menor valor *BIC*. Em seguida (linha 3), é construído o modelo definitivo, o qual agrupa o conjunto *D*, com o número de *clusters* recém descoberto. Em posse do modelo definitivo, a matriz de probabilidades Ω é computada a partir da Definição 11 (linha 4). A partir do *loop* que se inicia na linha 5, os graus de correlação entre quaisquer pares de indivíduos começam a ser computados para a formação da matriz Δ. A correlação entre dois indivíduos é mensurada a partir do produto entre seus respectivos coeficientes de probabilidade, o qual representa a probabilidade desses dois indivíduos pertencerem a *clusters* iguais, e coeficiente de correlação de *Spearman*. Por fim, na linha 11, a matriz Δ é retornada.

Nessa abordagem, o coeficiente de *Spearman* é computado entre todos os pares de indivíduos, assim como na solução apresentada pelo *baseline*. Entretanto, diferente do *baseline*, o qual aplica diretamente o coeficiente de correlação de *Spearman* entre dois indivíduos, considerando, a priori, que os mesmos já se correlacionam, nossa abordagem prioriza a probabilidade desses indivíduos se correlacionarem entre si, a fim de diminuir o impacto do coeficiente de *Spearman* no resultado final do grau de correlação.

Exposto o método, repetiremos os experimentos apresentados na abordagem anterior com o objetivo de avaliar a proposta desta segunda abordagem em relação à primeira. Contudo, não há necessidade do experimento que avalia o impacto do número de *clusters* sobre o ERM.

Algoritmo 4: ETAPA_1 (Abordagem 2)

Entrada: Conjunto de dados D
Saída: Matriz de graus de correlação Δ

início
  $numero\_de\_clusters \leftarrow \text{BIC}(D)$;
  $modelo \leftarrow \text{GMM}(D, numero\_de\_clusters)$;
  $\Omega \leftarrow \text{CONSTROI\_}\Omega(modelo, numero\_de\_clusters)$;
  `// Indivíduo i`
  para ***cada*** $i \in D$ **faça**
    `// Indivíduo j`
    para ***cada*** $j \in D$ **faça**
      $\delta_{ij} \leftarrow \Omega_{ij} \times \text{SPEARMAN}(i, j)$;
      $\Delta_{ij} \leftarrow \delta_{ij}$;
    fim
  fim
  retorna Δ;
fim

Esse experimento foi removido devido à existência do artifício *BIC*, o qual é empregado junto ao algoritmo *GMM* para descobrir o número de *clusters* que melhor agrupa um conjunto de dados específico. Portanto, não faria sentido estimar vários valores hipotéticos para os parâmetros de configuração do algoritmo *GMM*, como foi feito na abordagem 1 ao utilizar o algoritmo *DBSCAN*.

Tabela 10 – Número de *clusters* identificados através do artifício *BIC* para a construção do modelo, utilizando o algoritmo *GMM*, de cada conjunto de dados.

Conjunto de Dados	**Número de *Clusters***
Adult	60
CCC	32
SBRHAPT	79

No primeiro experimento, apresentado na Figura 14, avaliamos o impacto do parâmetro ε da Privacidade Diferencial sobre o erro relativo médio. Novamente, o parâmetro ε foi fixado em 1.0, enquanto o número de *clusters* do modelo *GMM* de cada conjunto de dados foi fixado conforme os valores especificados na Tabela 10. Os parâmetros de configuração do algoritmo *DBSCAN*, utilizados pela abordagem 1, permanecem inalterados em relação ao experimento 12. O gráfico mantém o comportamento esperado em relação ao ε, onde o erro relativo médio tende a aumentar de acordo com o decaimento de ε. Além disso, como é possível observar, os resultados provenientes da abordagem 2 se mostraram mais úteis em relação à abordagem 1. Isso

se deve ao artifício *BIC*, o qual nos permite construir um modelo mais preciso para representar o conjunto de dados, e ao coeficiente de probabilidade que foi inserido no cálculo da correlação, o qual permite realizar uma análise mais precisa dos relacionamentos entre os indivíduos.

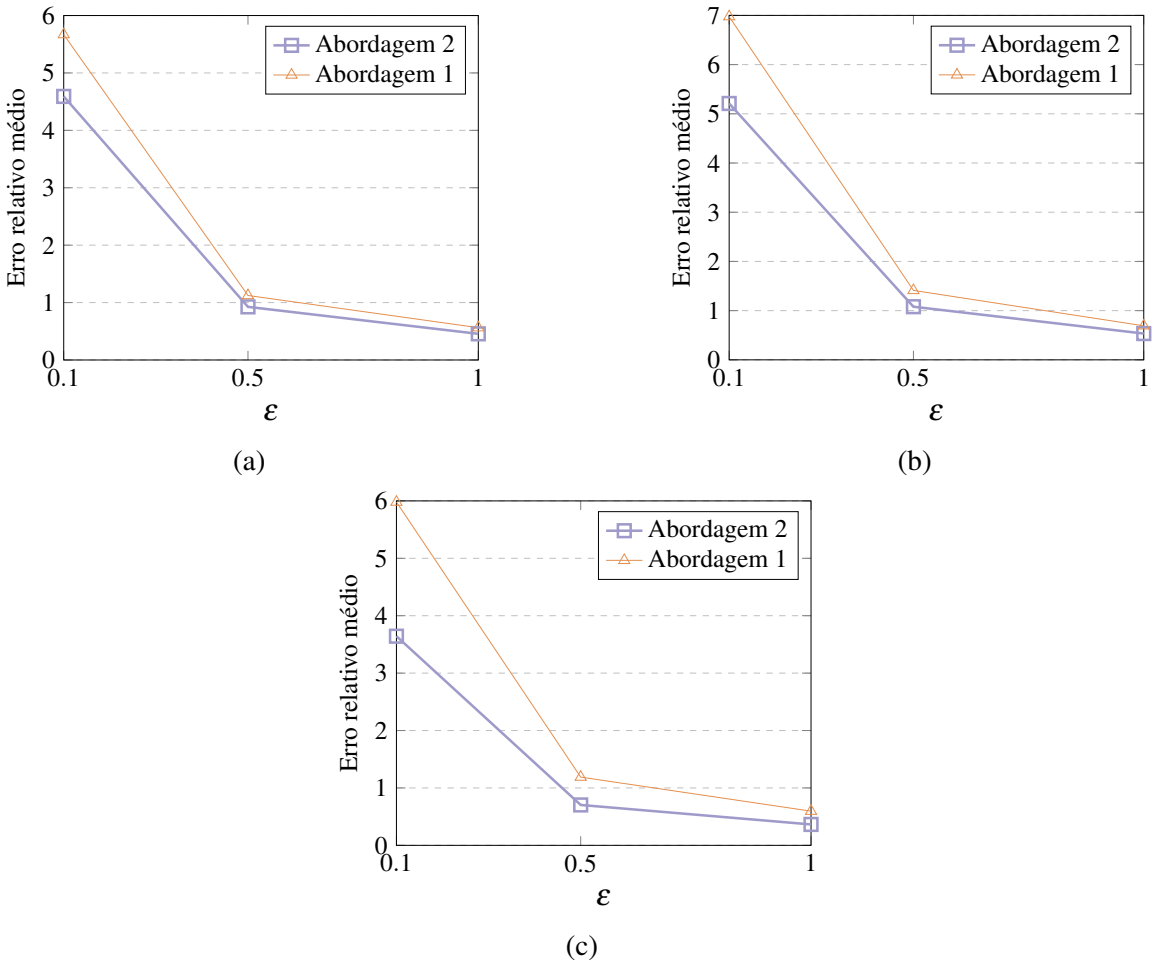


Figura 14 – Erro relativo resultante ao variar o parâmetro ε do modelo de Privacidade Diferencial. (a) *Adult*. (b) *CCC*. (c) *SBRHAPT*.

Por fim, finalizando os experimentos que avaliam a abordagem 2, a Figura 15 compara as respostas reais das consultas com suas versões anonimizadas provenientes dos métodos empregados nas abordagens 1 e 2. Como demonstrado nos gráficos, os resultados são uma consequência direta da Figura 14, a qual comprova que o erro relativo médio proveniente da abordagem 2 é menor em comparação à abordagem 1. Um erro relativo médio menor assegura que as respostas obtidas por meio de um mecanismo diferencialmente privado serão, em sua grande maioria, menores, resultando em uma maior utilidade. Visto que o erro relativo médio obtido pela abordagem 2 é menor, suas respostas anonimizadas se localizam mais próximas da diagonal que representa as respostas reais. Embora possa parecer que uma resposta anonimizada com uma melhor utilidade, ou seja, mais próxima de sua resposta real, resulte em uma perda

de privacidade, todas as propriedades do modelo de Privacidade Diferencial continuam sendo garantidas. Portanto, o método proposto não fere a privacidade de nenhum indivíduo.

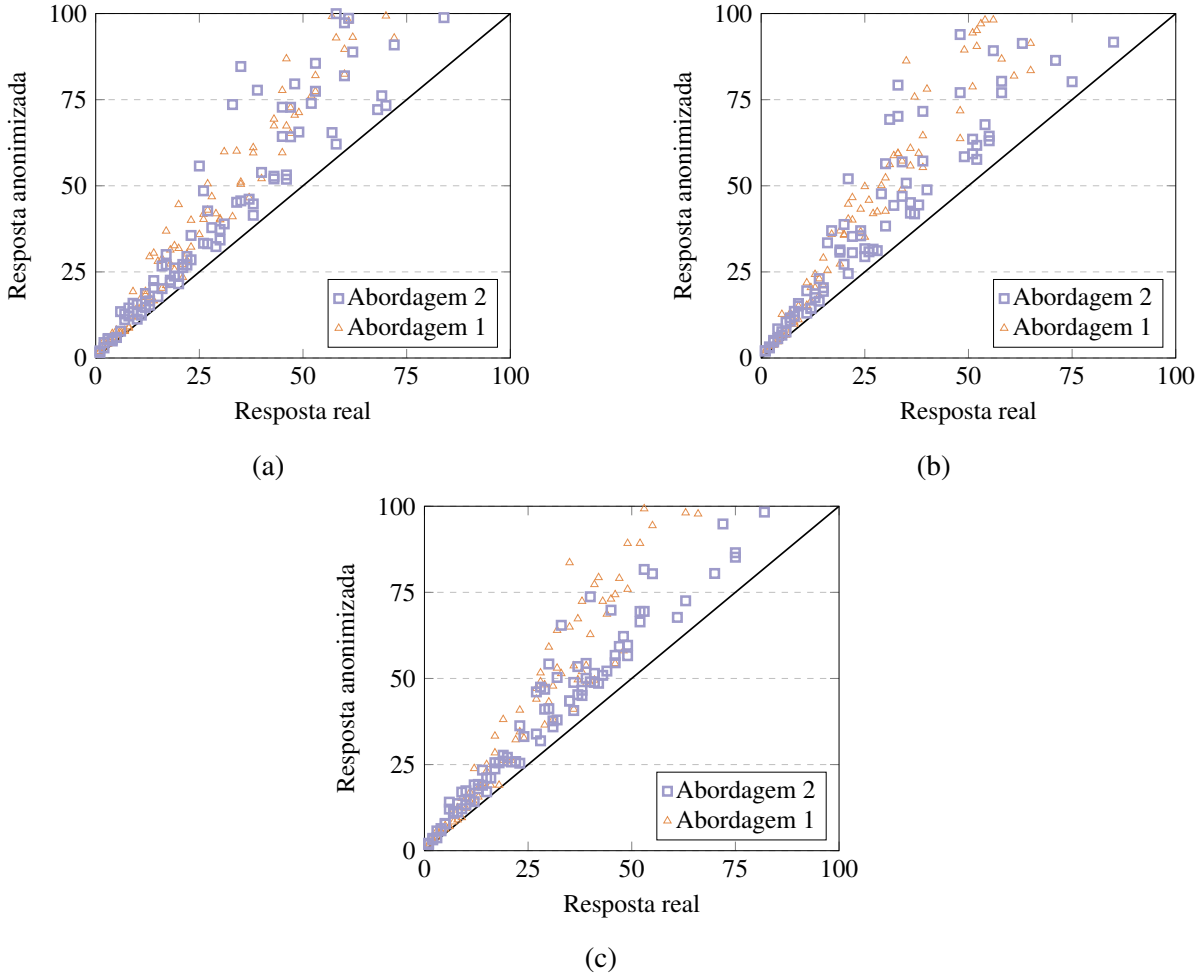


Figura 15 – Comparativo entre as respostas reais e suas versões anonimizadas. (a) *Adult*. (b) *CCC*. (c) *SBRHAPT*.

4.3.3 Abordagem 3: Construção da Matriz de Graus de Correlação com GMM e Redução de Dimensionalidade

A abordagem 3 consiste na última de nossas abordagens propostas com objetivo de solucionar o problema da preservação de privacidade em conjuntos de dados correlacionados através de técnicas de agrupamento. Como apresentado nas duas abordagens anteriores, abordagens 1 e 2, a utilidade dos dados obtidos por meio de um mecanismo diferencialmente privado é maior do que em trabalhos anteriores, ao mesmo tempo em que a privacidade dos indivíduos presentes no conjunto de dados permanece preservada. Entretanto, foi observado que o tempo decorrido para identificar os relacionamentos entre os indivíduos e mensurar seus respectivos graus de

correlação é uma tarefa que demanda bastante tempo. Isso torna as abordagens existentes, de certa forma, inviáveis, visto que, em ambientes interativos, onde os usuários requisitam consultas a um conjunto de dados e recebem uma resposta anonimizada, a requisição de um usuário deve ser atendida o mais breve possível. Portanto, essa abordagem tem o objetivo de melhorar o desempenho do mecanismo empregado, principalmente agilizando o processo de descoberta e computação da correlação entre os indivíduos, ao passo que a utilidade e a privacidade dos dados é preservada.

O método empregado na abordagem 3, a fim de melhorar o desempenho do mecanismo, consiste em utilizar uma técnica de seleção de atributos antes da etapa de agrupamento do conjunto de dados. Como algoritmo de agrupamento, o *GMM* continua sendo utilizado, visto que apresentou melhores resultados em relação ao algoritmo *DBSCAN*, além de possuir artifícios que permitem construir um melhor modelo. A etapa de computação dos graus de correlação entre indivíduos também foi modificada, de forma que o coeficiente de correlação de *Spearman* não é mais utilizado. O principal motivo de sua remoção é sua alta complexidade quando comparado aos graus de probabilidade. Computar o coeficiente de *Spearman* demanda muito mais tempo que os graus de probabilidade, os quais identificam quando dois indivíduos pertencem a *clusters* iguais. Além disso, utilizar os dois coeficientes se torna um tanto redundante, visto que suas propostas são similares. O coeficiente de *Spearman* afirma o quão similar dois objetos são entre si, enquanto o grau de probabilidade diz respeito à probabilidade de dois objetos pertencerem a *clusters* iguais. Portanto, uma vez que as técnicas de agrupamento são utilizadas para identificar objetos similares entre si, ambos os coeficientes podem ser utilizados para mensurar a correlação entre indivíduos.

Por sua vez, a seleção de atributos escolhe um subconjunto de atributos que representa, satisfatoriamente, o conjunto de dados em sua totalidade. Em nosso contexto, a utilização de uma técnica de seleção de atributos tem o objeto de simplificar o modelo gerado a partir do algoritmo *GMM*, reduzir a dimensionalidade dos dados e, por fim, reduzir o sobreajuste (*overfitting*).

A *Recursive Feature Elimination* / Eliminação Recursiva de Atributo (*RFE*) (GUYON *et al.*, 2002) foi a técnica de seleção de atributos utilizada para eliminar os atributos irrelevantes, ou redundantes, resultando em um conjunto de dados mais representativo e de menor tamanho, onde os atributos resultantes podem ser equiparados aos atributos semi-identificadores das abordagens sintáticas de privacidade de dados. O método *RFE* consiste em remover os atributos de

forma recursiva, determinando pesos para os atributos através de um estimador externo, como, por exemplo, *Random Forest* / Floresta Aleatória (*RF*) (GRANITTO *et al.*, 2006) e *Support Vector Machine* / Máquina de Vetor de Suporte (*SVM*) (HEARST *et al.*, 1998).

A recursividade do método é necessária devido à importância relativa de cada atributo, uma vez que seus valores podem variar substancialmente quando comparados a um subconjunto de atributos durante o processo de eliminação gradual, em particular para atributos com elevada correlação. Portanto, o *RFE* avalia o impacto de todos os subconjuntos de atributos.

Em nossa abordagem, escolhemos o método *RF* como estimador da técnica *RFE*, em oposição ao método *SVM*, uma vez que obtivemos melhores resultados associados a uma execução muito mais rápida. A grosso modo, o método *RF* utiliza um conjunto de árvores, onde cada árvore é instanciada com uma amostra do conjunto de dados de entrada e em cada divisão da árvore um subconjunto de atributos é escolhido. Após o treino de todas as árvores, o resultado é dado pelo voto da maioria das respostas provenientes de cada árvore preditora. O processo de instanciar cada árvore com amostras aleatórias diminui a correlação entre as árvores. Portanto, se um atributo é relevante para o conjunto resposta, este deverá estar presente em várias árvores.

A Figura 16 ilustra o funcionamento do método *RF*, enquanto a Tabela 11 apresenta o número de atributos removidos e restantes em cada conjunto de dados após a utilização do algoritmo *RFE* com o estimador *RF*.

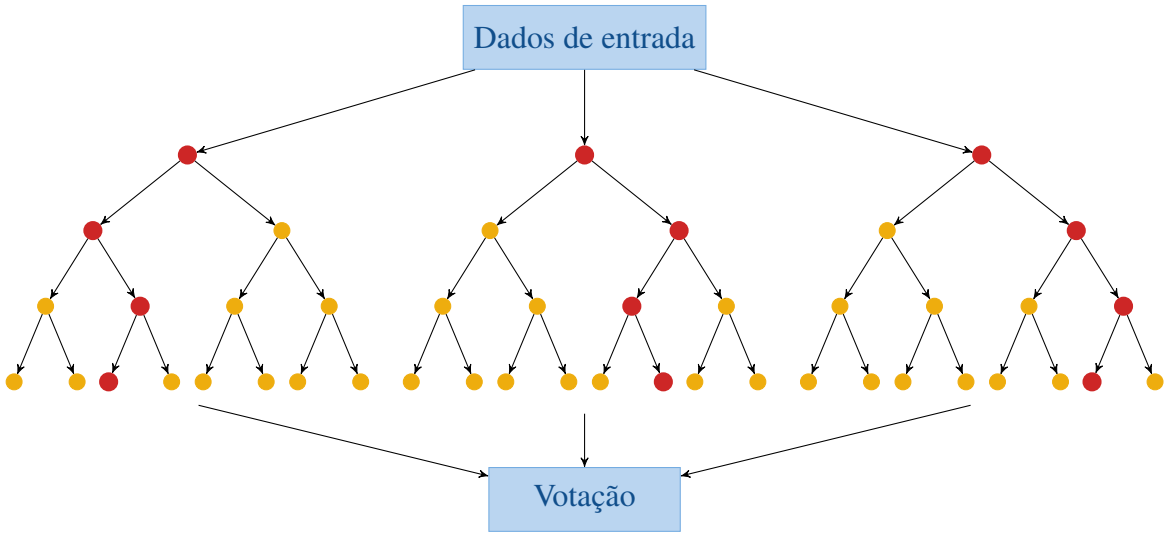


Figura 16 – Representação de uma Floresta Aleatória, composta por várias árvores de decisão que recebem um subconjunto de dados. A predição é dada em função da votação de cada árvore (CHAVES, 2017).

O Algoritmo 5 apresenta a solução proposta, a qual modifica a etapa 1 de nossa solução geral, utilizando o algoritmo *GMM* juntamente com uma técnica de seleção de atributos.

Tabela 11 – Número de atributos removidos e restantes em cada conjunto de dados após a seleção de atributos por meio do algoritmo *RFE*.

Conjunto de Dados	Número de Atributos Removidos	Número de Atributos Restantes
Adult	6	8
CCC	14	10
SBRHAPT	501	60

Além disso, a solução também provê mudanças na computação dos graus de correlação entre os indivíduos. O algoritmo é bastante semelhante ao Algoritmo 4 utilizado como solução para a abordagem 2. O parâmetros de entrada e saída permanecem inalterados, um conjunto de dados D e a matriz de graus de correlação Δ, respectivamente. As maiores mudanças ocorrem nas linhas 2 e 8. Na linha 2, o conjunto de entrada D é submetido a uma seleção de atributos por meio do algoritmo *RFE*, utilizando o método *RF* como estimador, transformando-se em um conjunto de dados modificado denominado D^*. O conjunto D^* passa a ser utilizado em todas as etapas consecutivas, como: identificação do número de *clusters* através do artifício *BIC*, construção do modelo *GMM* e construção da matriz de probabilidades Ω. A segunda mudança, ocorrida na linha 8, consiste na computação dos graus de correlação entre indivíduos, onde não há mais a participação do coeficiente de correlação de *Spearman*, apenas dos coeficientes de probabilidade.

Algoritmo 5: ETAPA_1 (Abordagem 3)

Entrada: Conjunto de dados D
Saída: Matriz de graus de correlação Δ

```
1  início
2      D* ← RFE(D);
3      numero_de_clusters ← BIC(D*);
4      modelo ← GMM(D*, numero_de_clusters);
5      Ω ← CONSTROI_Ω(modelo, numero_de_clusters);
       // Indivíduo i
6      para cada i ∈ D* faça
           // Indivíduo j
7          para cada j ∈ D* faça
8              δ_ij ← Ω_ij;
9              Δ_ij ← δ_ij;
10         fim
11     fim
12     retorna Δ;
13 fim
```

Os experimentos adiante finalizam o conjunto de experimentos realizados com o intuito de avaliar as abordagens propostas. Nos experimentos, avaliamos a abordagem 3, no que

Tabela 12 – Número de *clusters* identificados através do artifício *BIC* para a construção do modelo, utilizando o algoritmo *GMM*, de cada conjunto de dados após a seleção de atributos por meio do algoritmo *RFE*.

Conjunto de Dados	**Número de *Clusters***
Adult	12
CCC	9
SBRHAPT	76

diz respeito à utilidade e privacidade dos dados, em comparação à abordagem 2. Já a análise de desempenho avalia todas as nossas abordagens e o *baseline*.

O experimento da Figura 17 avalia o impacto do parâmetro ε da Privacidade Diferencial sobre o ERM e, assim como na avaliação das outras abordagens, o seu valor foi fixado em 1.0. Durante a execução desse experimento, os modelos provenientes do algoritmo *GMM* também foram fixados. Para isso, os conjuntos de dados com seus respectivos números de *clusters* foram definidos conforme especificado na Tabela 10, para a abordagem 2, e Tabela 12, para a abordagem 3. Como demonstrado nos gráficos, os resultados obtidos pela abordagem 3 ficaram levemente acima dos resultados obtidos por meio da abordagem 2, com exceção do experimento sobre o conjunto de dados *CCC*, Figura 17b, onde a abordagem apresentou melhores resultados para os valores de ε igual a 0.5 e 1.0. Isso ocorre devido à redundância do método empregado na abordagem 2, o qual utiliza dois coeficientes, com propostas semelhantes, para o cálculo da correlação. Além disso, visto que o coeficiente de correlação de *Spearman* pertence ao intervalo [-1, -1] e o coeficiente de probabilidade ao intervalo [0, 1], o produto de seus respectivos valores sempre ocasionará em resultados finais com valores ainda mais baixos. Apesar disso, o método utilizado na abordagem 3 se mostrou bastante eficaz, uma vez que foi empregado apenas o coeficiente de probabilidade como sendo o próprio coeficiente de correlação e, ainda assim, obteve resultados semelhantes e, até mesmo, melhores. A etapa de seleção de atributos contribuiu bastante para isso, uma vez que apenas os atributos semi-identificadores foram utilizados no cálculo da correlação, tornando os resultados mais precisos.

A Figura 18 consiste no último experimento que avalia a utilidade e privacidade. Nele, comparamos as respostas reais das consultas com suas versões anonimizadas providas pelos mecanismos das abordagens 2 e 3. Novamente, o parâmetro ε foi fixado em 1.0 durante a execução dos mecanismos. Como esperado, em consequência dos resultados obtidos no experimento anterior, representado pela Figura 17, o qual apresenta resultados similares para ambas as abordagens, as respostas anonimizadas providas por ambas as abordagens apresentam valores muito próximos uns dos outros. Além disso, por conta da aleatoriedade dos mecanismos,

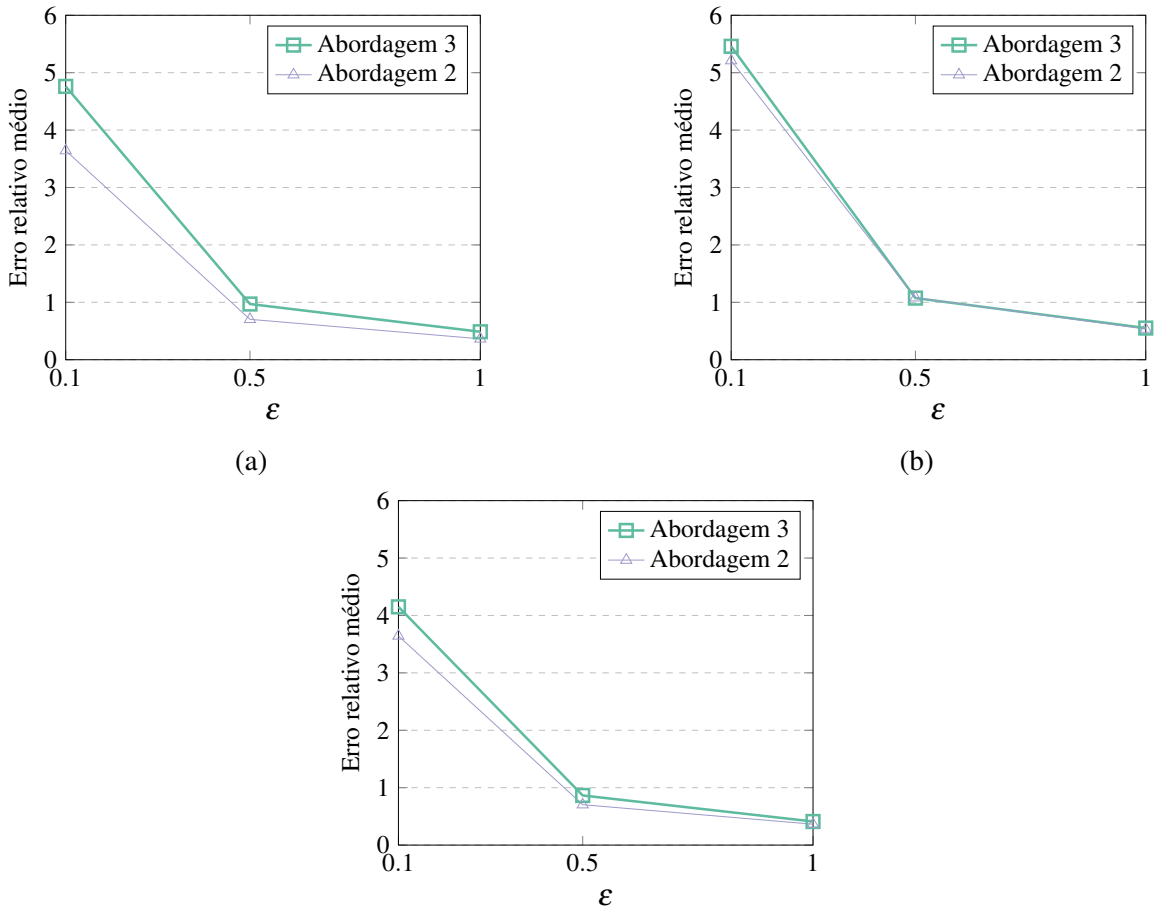


Figura 17 – Erro relativo resultante ao variar o parâmetro ε do modelo de Privacidade Diferencial. (a) *Adult*. (b) *CCC*. (c) *SBRHAPT*.

as respostas anonimizadas alternam entre si. Ora as respostas providas pelo mecanismo da abordagem 3 se posicionam mais próximas da diagonal que representa as respostas reais, ora são as respostas providas pelo mecanismo da abordagem 2. A única exceção ocorre no experimento executado sobre o conjunto de dados *CCC*, representado pela Figura 18b, o qual apresentou um erro relativo médio menor, conforme visto no experimento anterior, e, portanto, apresenta, em sua grande maioria, respostas anonimizadas mais próximas das respostas reais e abaixo dos resultados providos pela abordagem 2.

Finalizando a avaliação experimental, apresentamos na Figura 19 uma análise de desempenho das nossas abordagens e do *baseline*. A análise contabiliza apenas o tempo necessário, em minutos, para a construção da matriz de graus de correlação Δ. O tempo decorrido durante a execução do mecanismo foi desconsidera, a qual inclui calcular as sensibilidades das consultas e retornar suas respostas anonimizadas, visto que o tempo é igual em todas as abordagens. Além disso, sempre que acrescentássemos mais consultas em nossa configuração

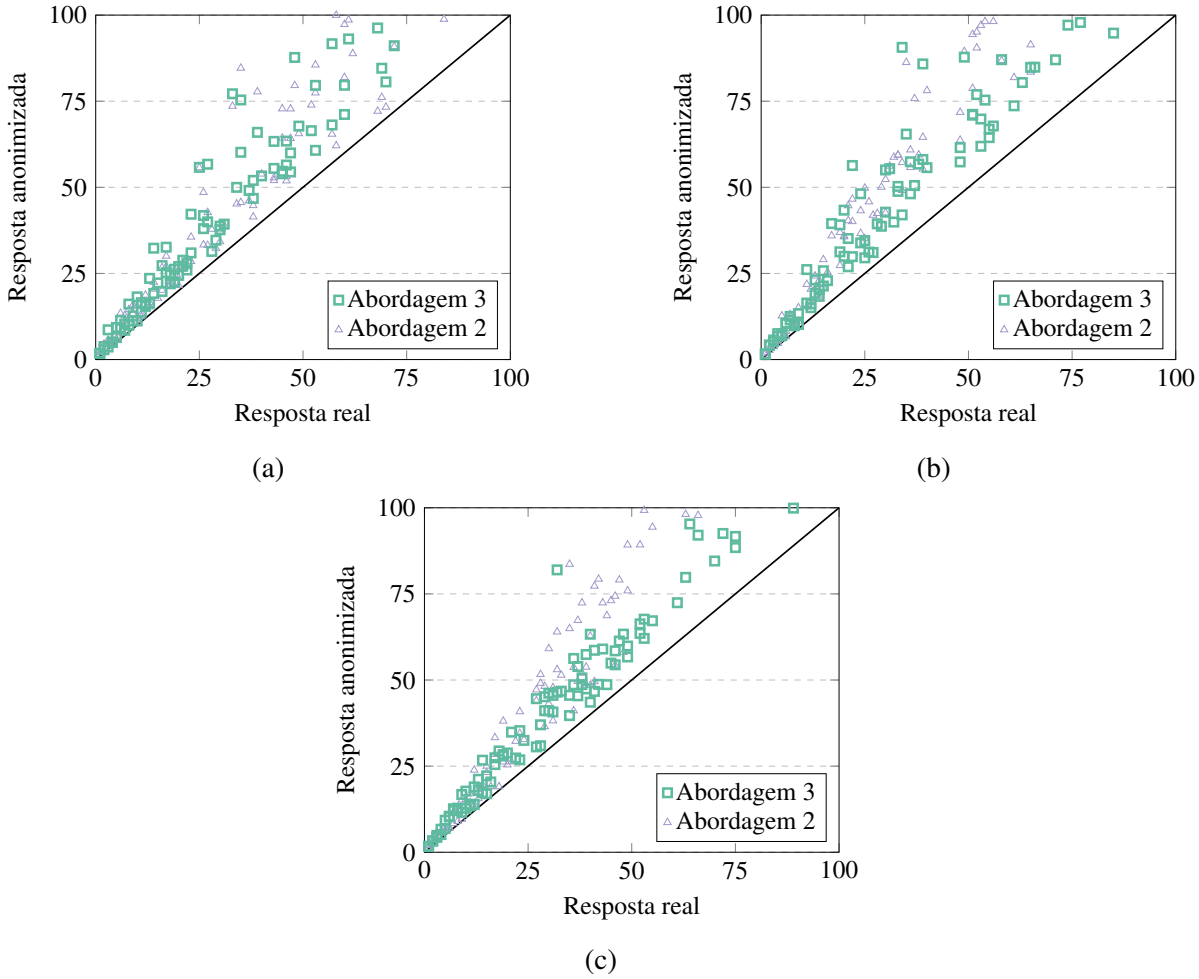


Figura 18 – Comparativo entre as respostas reais e suas versões anonimizadas. (a) *Adult*. (b) *CCC*. (c) *SBRHAPT*.

experimental, o tempo empenhado na execução do mecanismo seria ainda maior, ou seja, o tempo é proporcional ao número de consultas, e poderia distorcer o gráfico. Por fim, é importante ressaltar que a etapa de construção da matriz Δ é a mais custosa dentro das abordagens. Todos os métodos empregados, em suas respectivas abordagens, tais como: coeficiente de correlação de *Spearman*, *BIC*, construção de matriz de probabilidades Ω e *RFE*, tiveram seus tempos contabilizados.

Como apresentado no gráfico da Figura 19, o tempo decorrido para a construção da matriz Δ através dos métodos empregados pelo *baseline*, abordagens 1 e 2 são relativamente semelhantes. A pequena diferença apresentada entre o *baseline* e a abordagem 2 ocorre devido à necessidade de ambas calcularem o coeficiente de correlação de *Spearman* entre todos os pares de indivíduos. Além disso, na abordagem 2, ainda existe a necessidade de criar a matriz de probabilidades Ω. No entanto, a complexidade do coeficiente de *Spearman* é muito maior que o coeficiente de probabilidade, demandando muito mais tempo para ser computada. Isso

é comprovado pelos resultados obtidos pela abordagem 3, a qual utiliza apenas os coeficientes de probabilidade como os próprios graus de correlação entre os indivíduos. Por sua vez, a abordagem 2, a qual utiliza o algoritmo *DBSCAN*, apresenta tempos intermediários entre o *baseline* e a abordagem 2, visto que nela são contabilizadas apenas as correlações entre os indivíduos pertencentes aos mesmos *clusters*, e não entre todos os pares possíveis. Além de excelentes resultados, a abordagem 3 apresentou expressivos ganhos de desempenho em relação às outras abordagens mencionadas ao longo do capítulo, sendo, aproximadamente, 40x mais rápida sobre os conjuntos de dados *Adult* e *CCC* e 8x mais rápida sobre o conjunto de dados *SBRHAPT*.

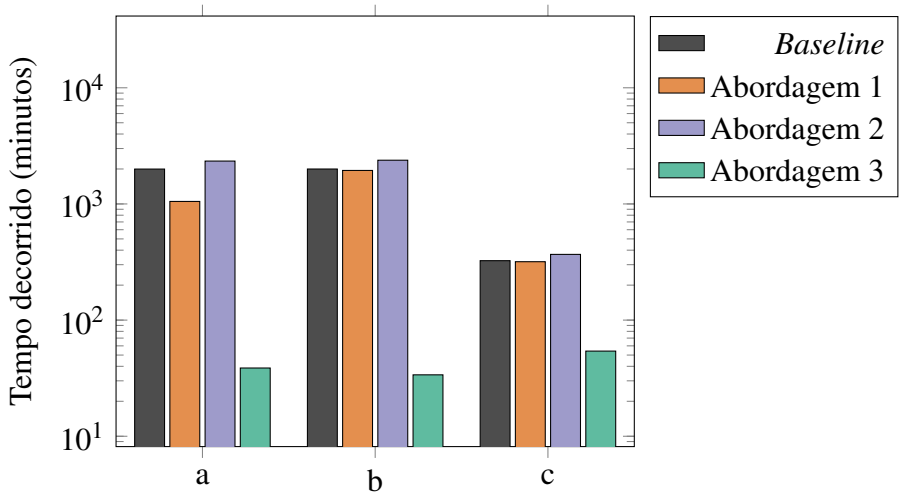


Figura 19 – Comparativo de desempenho. (a) *Adult*. (b) *CCC*. (c) *SBRHAPT*.

4.4 Conclusão

Neste capítulo foram apresentadas as abordagens propostas para a solução do problema de preservação de privacidade em conjunto de dados correlacionados, especificamente através do modelo de Privacidade Diferencial. Além disso, também foi apresentada uma análise experimental de nossas abordagens, juntamente com o *baseline*, utilizando conjuntos de dados reais. As abordagens propostas utilizam técnicas de agrupamento com o objetivo de identificar os relacionamentos existentes em um conjunto de dados e mensurá-los posteriormente. A solução da abordagem 3, a mais promissora, emprega, ainda, uma técnica de seleção de atributos para tornar a computação dos relacionamentos mais precisa e menos custosa. Experimentos variando uma série de parâmetros também demonstram a eficiência de nossas abordagens, tanto em termos de utilidade quanto em desempenho.

5 CONSIDERAÇÕES FINAIS

5.1 Conclusão

Neste trabalho, apresentamos três abordagens com soluções distintas para o problema de preservação de privacidade em conjuntos de dados correlacionados em ambientes interativos. Para garantir a efetividade de nossas soluções, utilizamos técnicas de agrupamento e o modelo de Privacidade Diferencial, de forma que um usuário malicioso não seja capaz de realizar um ataque probabilístico e ferir a privacidades dos indivíduos.

As abordagens são divididas em duas etapas. A primeira etapa consiste em identificar os relacionamentos existentes entre os indivíduos, através de técnicas de agrupamento, e mensurar seus respectivos graus de correlação por meio de uma medida de correlação. Em nosso contexto, indivíduos se relacionam entre si quando estes pertencem ao mesmo *cluster*. Inicialmente, utilizamos o algoritmo *DBSCAN* como técnica de agrupamento, sendo, posteriormente, substituído pelo algoritmo *GMM*. O coeficiente de correlação de postos de *Spearman* foi empregado, inicialmente, como a medida de correlação. Entretanto, devido à sua alta complexidade, foi substituído pelo coeficiente de probabilidade, o qual representa a probabilidade de dois indivíduos pertencerem ao mesmo *clusters* e, portanto, se correlacionarem entre si. Além disso, ambos os coeficientes apresentam propostas semelhantes. Por fim, utilizamos uma técnica de seleção de atributos com o intuito de manter apenas os atributos semi-identificadores e tornar a etapa de agrupamento mais precisa, melhorando a utilidade dos dados. A segunda etapa consiste em aplicar um mecanismo diferencialmente privado, considerando os relacionamentos computados na etapa anterior, a fim de garantir as propriedades do modelo de Privacidade Diferencial e preservar a privacidade dos indivíduos. Para tal, utilizamos o mecanismo de *Laplace* para prover informações anonimizadas.

Comparamos as abordagens com a estratégia proposta por Zhu et al. (ZHU *et al.*, 2015), a qual denominamos *baseline*. Foram realizados experimentos utilizando três conjuntos de dados reais com o objetivo de comprovar a qualidade das três soluções propostas em termos de utilidade e desempenho. Para isso, variamos os parâmetros de configuração dos algoritmos de agrupamento. No algoritmo *DBSCAN* variamos os parâmetros *eps* e $minPoints$, enquanto no algoritmo *GMM* apenas o número de *clusters*. Além disso, também variamos o parâmetro ε do modelo de Privacidade Diferencial. Os resultados obtidos em todos experimentos comprovam que nossas soluções são significativamente melhores em utilidade, independente do conjunto

de dados empregado. Quanto ao desempenho, destaca-se a nossa última solução proposta, que obteve expressivos ganhos em desempenho, mostrando-se até 40 vezes mais rápida que as outras soluções, inclusive em relação ao *baseline*.

5.2 Trabalhos Futuros

Como trabalhos futuros, pretendemos, inicialmente, utilizar alguma técnica de *data imputation* para estimar os atributos faltantes de alguns indivíduos, visto que isso ainda é uma limitação existente neste trabalho. Pretendemos, também, adaptar nossas soluções para um contexto online de *streaming* de dados. Por fim, como intenção mais ambiciosa, temos interesse em propor um novo mecanismo de Privacidade Diferencial, específico para o contexto de dados correlacionados, mais robusto e eficiente que os mecanismos presentes atualmente na literatura.

REFERÊNCIAS

BISHOP, C. M. **Pattern Recognition and Machine Learning (Information Science and Statistics)**. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006. ISBN 0387310738.

CHAVES, I. C. **Redes bayesianas para previsão de falhas em discos rígidos**. [S.l.]: Universidade Federal do Ceará, 2017.

CHEN, R.; FUNG, B. C.; PHILIP, S. Y.; DESAI, B. C. Correlated network data publication via differential privacy. **The VLDB Journal**, Springer, v. 23, n. 4, p. 653–676, 2014.

CLIFTON, C.; TASSA, T. On syntactic anonymity and differential privacy. **Transactions on Data Privacy**, v. 6, n. 2, p. 161–183, 2013.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm. **Journal of the royal statistical society. Series B (methodological)**, JSTOR, p. 1–38, 1977.

DEWRI, R.; RAY, I.; RAY, I.; WHITLEY, D. On the optimal selection of k in the k-anonymity problem. In: **24th ICDE International Conference on Data Engineering**. [S.l.: s.n.], 2008. p. 1364–1366.

DOMINGO-FERRER, J.; SÁNCHEZ, D.; SORIA-COMAS, J. Database anonymization: Privacy models, data utility, and microaggregation-based inter-model connections. **Synthesis Lectures on Information Security, Privacy, & Trust**, Morgan & Claypool Publishers, v. 8, n. 1, p. 1–136, 2016.

DWORK, C. Differential privacy. In: **33rd International Colloquium on Automata, Languages and Programming**. [S.l.: s.n.], 2006. p. 1–12.

DWORK, C. Differential privacy: A survey of results. In: SPRINGER. **International Conference on Theory and Applications of Models of Computation**. [S.l.], 2008. p. 1–19.

DWORK, C.; ROTH, A. *et al.* The algorithmic foundations of differential privacy. **Foundations and Trends® in Theoretical Computer Science**, Now Publishers, Inc., v. 9, n. 3–4, p. 211–407, 2014.

ESTER, M.; KRIEGEL, H.-P.; SANDER, J.; XU, X. *et al.* A density-based algorithm for discovering clusters in large spatial databases with noise. In: **KDD Knowledge Discovery and Data Mining**. [S.l.: s.n.], 1996. v. 96, n. 34, p. 226–231.

FUNG, B.; WANG, K.; CHEN, R.; YU, P. S. Privacy-preserving data publishing: A survey of recent developments. **ACM Computing Surveys (CSUR)**, ACM, v. 42, n. 4, p. 14, 2010.

GORDON, A. D. Classification, (chapman & hall/crc monographs on statistics & applied probability). Chapman and Hall/CRC, 1999.

GRANITTO, P. M.; FURLANELLO, C.; BIASIOLI, F.; GASPERI, F. Recursive feature elimination with random forest for ptr-ms analysis of agroindustrial products. **Chemometrics and Intelligent Laboratory Systems**, Elsevier, v. 83, n. 2, p. 83–90, 2006.

GUYON, I.; WESTON, J.; BARNHILL, S.; VAPNIK, V. Gene selection for cancer classification using support vector machines. **Machine learning**, Springer, v. 46, n. 1-3, p. 389–422, 2002.

HEARST, M. A.; DUMAIS, S. T.; OSUNA, E.; PLATT, J.; SCHOLKOPF, B. Support vector machines. **IEEE Intelligent Systems and their applications**, IEEE, v. 13, n. 4, p. 18–28, 1998.

JONES, E.; OLIPHANT, T.; PETERSON, P. *et al.* **SciPy: Open source scientific tools for Python**. 2001. [Online; accessed <today>]. Disponível em: <http://www.scipy.org/>.

KIFER, D.; MACHANAVAJJHALA, A. No free lunch in data privacy. In: ACM. **Proceedings of the 2011 ACM SIGMOD International Conference on Management of data**. [S.l.], 2011. p. 193–204.

KIFER, D.; MACHANAVAJJHALA, A. Pufferfish: A framework for mathematical privacy definitions. **ACM Transactions on Database Systems (TODS)**, ACM, v. 39, n. 1, p. 3, 2014.

LEE, J.; CLIFTON, C. How much is enough? choosing ε for differential privacy. In: SPRINGER. **International Conference on Information Security**. [S.l.], 2011. p. 325–340.

LI, N.; LI, T.; VENKATASUBRAMANIAN, S. t-closeness: Privacy beyond k-anonymity and l-diversity. In: **23th ICDE International Conference on Data Engineering (ICDE)**. [S.l.: s.n.], 2007. p. 106–115.

LICHMAN, M. **UCI Machine Learning Repository**. 2013. Disponível em: <http://archive.ics.uci.edu/ml>.

LINDEN, R. Técnicas de agrupamento. **Revista de Sistemas de Informação da FSMA**, v. 4, p. 18–36, 2009.

LIU, C.; CHAKRABORTY, S.; MITTAL, P. Dependence makes you vulnerable: Differential privacy under dependent tuples. In: **NDSS Network and Distributed System Security Symposium**. [S.l.: s.n.], 2016.

MACHANAVAJJHALA, A.; GEHRKE, J.; KIFER, D.; VENKITASUBRAMANIAM, M. l-diversity: Privacy beyond k-anonymity. p. 24–24, 2006.

MCKINNEY, W. Data structures for statistical computing in python. In: WALT, S. van der; MILLMAN, J. (Ed.). **Proceedings of the 9th Python in Science Conference**. [S.l.: s.n.], 2010. p. 51 – 56.

MENDONÇA, A. L.; BRITO, F. T.; LINHARES, L. S.; MACHADO, J. C. Dipcoding: A differentially private approach for correlated data with clustering. In: ACM. **Proceedings of the 21st International Database Engineering & Applications Symposium**. [S.l.], 2017. p. 291–297.

MEYERSON, A.; WILLIAMS, R. On the complexity of optimal k-anonymity. In: **Proceedings of the 23rd ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Paris, France**. [S.l.: s.n.], 2004. p. 223–228.

NERGIZ, M. E.; ATZORI, M.; CLIFTON, C. Hiding the presence of individuals from shared databases. In: **Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data**. New York, NY, USA: ACM, 2007. (SIGMOD '07), p. 665–676. ISBN 978-1-59593-686-8. Disponível em: <http://doi.acm.org/10.1145/1247480.1247554>.

PEARSON, K. Note on regression and inheritance in the case of two parents. **Proceedings of the Royal Society of London**, JSTOR, v. 58, p. 240–242, 1895.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PFEIFFER, P. **Concepts of Probability Theory**. Dover Publications, 1978. (Dover Books on Mathematics). ISBN 9780486636771. Disponível em: <https://books.google.com.br/books?id=_mayRBczVRwC>.

SÁNCHEZ, D.; DOMINGO-FERRER, J.; MARTÍNEZ, S.; SORIA-COMAS, J. Utility-preserving differentially private data releases via individual ranking microaggregation. **Information Fusion**, Elsevier, v. 30, p. 1–14, 2016.

SCHOLZ, F. Maximum likelihood estimation. **Encyclopedia of statistical sciences**, Wiley Online Library, 1985.

SCHWARZ, G. *et al.* Estimating the dimension of a model. **The annals of statistics**, Institute of Mathematical Statistics, v. 6, n. 2, p. 461–464, 1978.

SHOARAN, M.; THOMO, A.; WEBER, J. H. Differential privacy in practice. In: **Secure Data Management - 9th VLDB Workshop, SDM 2012, Istanbul, Turkey, August 27, 2012. Proceedings**. [S.l.: s.n.], 2012. p. 14–24.

SORIA-COMAS, J.; DOMINGO-FERRER, J.; SÁNCHEZ, D.; MARTÍNEZ, S. Enhancing data utility in differential privacy via microaggregation-based k-anonymity. **The VLDB Journal**, Springer, v. 23, n. 5, p. 771–794, 2014.

SPEARMAN, C. The proof and measurement of association between two things. **The American journal of psychology**, JSTOR, v. 15, n. 1, p. 72–101, 1904.

SWEENEY, L. k-anonymity: A model for protecting privacy. **International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems**, World Scientific, v. 10, n. 05, p. 557–570, 2002.

THE Data Mining Hypertextbook. 2017. <http://www.hypertextbookshop.com/dataminingbook/public_version/contents/chapters/chapter004/section004/blue/page003.html>. Acessado em 10 de Dezembro, 2017.

WILLISON, D. J.; EMERSON, C.; SZALA-MENEOK, K. V.; GIBSON, E.; SCHWARTZ, L.; WEISBAUM, K. M.; FOURNIER, F.; BRAZIL, K.; COUGHLIN, M. D. Access to medical records for research purposes: varying perceptions across research ethics boards. **Journal of Medical Ethics**, Institute of Medical Ethics, v. 34, n. 4, p. 308–314, 2008.

XIAO, X.; WANG, G.; GEHRKE, J. Differential privacy via wavelet transforms. **IEEE Trans. Knowl. Data Eng.**, v. 23, n. 8, p. 1200–1214, 2011.

ZHU, T.; XIONG, P.; LI, G.; ZHOU, W. Correlated differential privacy: hiding information in non-iid data set. **IEEE Transactions on Information Forensics and Security**, IEEE, v. 10, n. 2, p. 229–242, 2015.

ZHU, Y. **Yu's Machine Learning Garden**. 2014. Disponível em: <http://yulearning.blogspot.com.br/2014/11/einsteins-most-famous-equation-is-emc2.html>.