Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Pareamento Privado de Atributos no Contexto da Resolução de Entidades com Preservação de Privacidade

Thiago Pereira da Nóbrega

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande - Campus I como parte dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Ciência da Computação
Linha de Pesquisa: Sistemas de Informação e Banco de Dados

Prof. Dr. Carlos Eduardo Santos Pires
(Orientador)

Campina Grande, Paraíba, Brasil

N754p Nóbrega, Thiago Pereira da.
Pareamento privado de atributos no contexto da resolução de entidades com preservação de privacidade / Thiago Pereira da Nóbrega. ï Campina Grande, 2018.
86 f. : il. color.

Dissertação (Mestrado em Ciência da Computação) ï Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2018.
"Orientação: Prof. Dr. Carlos Eduardo Santos Pires".
Referências.

1. Segurança e Privacidade. 2. Resolução de Entidades. 3. Integração de Dados. 4. Schema Matching. I. Pires, Carlos Eduardo Santos. II. Título.

CDU 004.056.53(043)

FICHA CATALOGRÁFICA ELABORADA PELO BIBLIOTECÁRIO GUSTAVO DINIZ DO NASCIMENTO CRB - 15/515

"PAREAMENTO PRIVADO DE ATRIBUTOS NO CONTEXTO DA RESOLUÇÃO DE ENTIDADES COM PRESERVAÇÃO DE PRIVACIDADE"

THIAGO PEREIRA DA NÓBREGA

DISSERTAÇÃO APROVADA EM 11/05/2018

CARLOS EDUARDO SANTOS PIRES, Dr., UFCG
Orientador(a)

CLÁUDIO ELÍZIO CALAZANS CAMPELO, PhD., UFCG
Examinador(a)

VALERIA CESARIO TIMES, Ph.D, UFPE
Examinador(a)

CAMPINA GRANDE - PB

Resumo

A Resolução de Entidades com Preservação de Privacidade (REPP) consiste em identificar entidades (e.g. Pacientes), armazenadas em bases de dados distintas, que correspondam a um mesmo objeto do mundo real. Como as entidades em questão possuem dados privados (ou seja, dados que não podem ser divulgados) é fundamental que a tarefa de REPP seja executada sem que nenhuma informação das entidades seja revelada entre os participantes (proprietários das bases de dados), de modo que a privacidade dos dados seja preservada. Ao final da tarefa de REPP, cada participante identifica quais entidades de sua base de dados estão presentes nas bases de dados dos demais participantes. Antes de iniciar a tarefa de REPP os participantes devem concordar em relação à entidade (em comum), a ser considerada na tarefa, e aos atributos das entidades a serem utilizados para comparar as entidades. Em geral, isso exige que os participantes tenham que expor os esquemas de suas bases de dados, compartilhando (meta-)informações que podem ser utilizadas para quebrar a privacidade dos dados. Este trabalho propõe uma abordagem semiautomática para identificação de atributos similares (pareamento de atributos) a serem utilizados para comparar entidades durante a REPP. A abordagem é inserida em uma etapa preliminar da REPP (etapa de Apresentação) e seu resultado (atributos similares) pode ser utilizado pelas etapas subsequentes (Blocagem e Comparação). Na abordagem proposta a identificação dos atributos similares é realizada utilizando-se representações dos atributos (Assinaturas de Dados), geradas por cada participante, eliminando a necessidade de divulgar informações sobre seus esquemas, ou seja, melhorando a segurança e privacidade da tarefa de REPP. A avaliação da abordagem aponta que a qualidade do pareamento de atributos é equivalente a uma solução que não considera a privacidade dos dados, e que a abordagem é capaz de preservar a privacidade dos dados.

Palavras-chave: Segurança e Privacidade, Resolução de Entidades, Integração de Dados, Schema Matching.

Abstract

The Privacy Preserve Record Linkage (PPRL) aims to identify entities, that can not have their information disclosed (e.g., Medical Records), which correspond to the same real-world object across different databases. It is crucial to the PPRL tasks that it is executed without revealing any information between the participants (database owners) during the PPRL task, to preserve the privacy of the original data. At the end of a PPRL task, each participant identifies which entities in its database are present in the databases of the other participants. Thus, before starting the PPRL task, the participants must agree on the entity and its attributes, to be compared in the task. In general, this agreement requires that participants have to expose their schemas, sharing (meta-)information that can be used to break the privacy of the data. This work proposes a semiautomatic approach to identify similar attributes (attribute pairing) to identify the entities attributes. The approach is inserted as a preliminary step of the PPRL (Handshake), and its result (similar attributes) can be used by subsequent steps (Blocking and Comparison). In the proposed approach, the participants generate a privacy-preserving representation (Data Signatures) of the attributes values that are sent to a trusted third-party to identify similar attributes from different data sources. Thus, by eliminating the need to share information about their schemas, consequently, improving the security and privacy of the PPRL task. The evaluation of the approach points out that the quality of attribute pairing is equivalent to a solution that does not consider data privacy, and is capable of preserving data privacy.

Keywords: Security and privacy, Entity Resolution, Data integration, Schema Matching

Agradecimentos

Ao meu orientador (Professor Carlos Eduardo), pela paciência e dedicação nas incontáveis revisões e orientações durante todo o mestrado. A minha esposa (Lêda) e minha filha (Maria Beatriz) por todo o apoio, incentivo, carinho e paciência que tiveram durante esses anos. Aos meus pais (Crizeuda e Paulo), por me ensinarem o valor do estudo. A minha irmã (Marcella), pelos momentos de alegria durante nossas vidas. Aos meus amigos do Laboratório de Qualidade de Dados (Demas, Brasileiro e Dimas), pelas discussões, revisões e pela companhia durante o mestrado. Aos amigos da UEPB (Bruno, Dannylo, Erick, W e Francinaldo) pela compreensão e apoio.

Conteúdo

Lista de Símbolos

RE - *Resolução de Entidades*

REPP - *Resolução de Entidades com Preservação de Privacidade*

QIDS - *quasi-identifiers*

MD5 - *Message Digest 5*

SHA - *Secure Hash Algorithm*

CH - *Criptografia Homomórfica*

FB - *Filtro de Bloom*

HPC - *Honesto porém curioso*

PAC - *Pareamento de Atributo às Cegas*

TPC - *Terceira Parte Confíavel*

AEI - *Assinatura de Equivalência da Informação*

ASD - *Assinatura de Similaridade dos Dados*

PPA - *Pareamento de Privado de Atributos*

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

As organizações (governos e empresas privadas) coletam e processam grandes conjuntos de dados para extrair conhecimento e auxiliar em tomadas de decisão. Com frequência, essas organizações necessitam combinar as informações de múltiplas bases de dados, oferecendo visões integradas do conjunto de dados com o intuito de aumentar a qualidade ou enriquecer a informação [13].

A integração de dados pode ser aplicada em diversas áreas de conhecimento: *Population Informatics*[1], Saúde, Segurança e Ciências Sociais [13, 39, 70]. Na área de Saúde tem-se o Google Flu and Dengue Trends [27] que integra dados médicos (fornecidos por hospitais) e informações de pesquisas sobre sintomas de doenças realizadas no buscador do Google com o objetivo de predizer surtos de dengue e gripe, de maneira mais precisa e rápida do que os métodos utilizados pelo Centro de Controle de Doença dos Estados Unidos [26]. Outro exemplo da utilização da integração de dados é o programa Longitudinal Employment Household Dynamics (LEHD) do Census Bureau dos Estados Unidos [19, 49, 56]. O LEHD integra dados de saúde, trabalho e educação de órgãos federais e dos 50 estados americanos. Este programa possibilitou que economistas desenvolvessem teorias e modelos em dados reais da economia norte-americana, com destaque para o trabalho sobre desemprego friccional que recebeu o prêmio Nobel de economia de 2010 [51].

[1]Population informatics é uma área do conhecimento que une Ciência Social, Medicina, Ciência da Computação e Estatística, e utiliza métodos quantitativos e ferramentas computacionais para responder a perguntas sobre populações humanas e grupos de indivíduos, aumentando o conhecimento sobre sociedade, saúde e comportamento humano [39].

Conforme ilustrado nos exemplos anteriores, a integração de dados possui um impacto importante em diversas áreas do conhecimento. Entretanto, para que os dados possam ser integrados e, consequentemente, gerar conhecimento, faz-se necessária a utilização de uma tarefa chamada Resolução de Entidades (RE). A RE visa identificar entidades (e.g., pessoas, restaurantes, publicações, produtos, entre outros) armazenadas em bases de dados distintas, que se referem a um mesmo objeto do mundo real [13]. Para identificar as entidades em comum armazenadas nas bases de dados, a RE compara as entidades (aos pares) aplicando funções de similaridade, as quais computam o grau de similaridade entre duas entidades com base nos dados de seus atributos.

A RE é relevante para diversas aplicações como, por exemplo, para aplicações que auxiliam na tomada de decisão, pois a existência de entidades duplicadas pode influenciar negativamente na interpretação das análises de dados. Por exemplo, considere o processo de aquisição de medicamentos e vacinas, no qual o estoque de todas as unidades de saúde é avaliado. Tal processo poderá ser comprometido com gastos desnecessários, caso as entidades (medicamentos e vacinas) duplicadas não sejam identificadas corretamente.

No contexto da tomada de decisão, a baixa qualidade dos dados influencia negativamente na interpretação das análises realizadas a partir destes dados e, consequentemente, compromete as decisões tomadas. Por exemplo, um processo de planejamento de uma cadeia de produção (envolvendo compra e estoque de matéria prima, produção e armazenamento de produtos) será muito provavelmente prejudicado caso sejam tomadas decisões baseadas em relatórios que contemplam dados de vendas com baixa acurácia, informações incompletas de fornecedores e/ou dados duplicados sobre o estoque de produtos.

Um problema recorrente que a RE enfrenta é a ausência de atributos capazes de identificar unicamente as entidades nas diferentes bases de dados. Isso impossibilita a utilização de operações de comparação simples (por exemplo, junções SQL), tornando necessária a realização de comparações sofisticadas envolvendo um conjunto de atributos comuns a todas as entidades. Tal conjunto de atributos é chamado de quasi-identifiers (QIDs) [13].

Diversos sistemas de informação coletam e armazenam dados nos quais aspectos como privacidade[2] e confidencialidade[3] devem ser considerados. São exemplos desses dados:

[2]Privacidade descreve o direito do indivíduo de controlar o acesso as suas informações pessoais, definindo como e quais informações outros indivíduos podem obter sobre si [4].

[3]Confidencialidade (ou sigilo) refere-se à forma como a informação privada, fornecida pelos indivíduos,

prontuários médicos, dados de navegação (GPS), transações financeiras, entre outros. Nesse contexto surge a Resolução de Entidades com Preservação de Privacidade (REPP), uma tarefa que visa integrar dados privados assegurando que a privacidade e a confidencialidade dos dados sejam preservadas durante toda a tarefa. Em outras palavras, além dos problemas intrínsecos à tarefa de RE (a citar, eficiência e eficácia da resolução), a REPP deve considerar a privacidade dos dados em todos os seus estágios, desde a escolha dos atributos a serem utilizados pela REPP, passando pela comparação entre as entidades, até a divulgação das entidades duplicadas presentes em cada base de dados (resultado da REPP).

1.1 Motivação/Problematização

No contexto da REPP, trabalhos de revisão da literatura (surveys) [38, 70] mostram que as abordagens do estado da arte exigem que os participantes (e.g., empresas, governos, entre outros) exponham algumas informações sobre as entidades que serão utilizadas na REPP. As informações reveladas pelos participantes da REPP normalmente incluem nomes, ordem e tipo de dado (inteiro, textual, ponto flutuante, data, entre outros) dos atributos que compõem as entidades. Essas informações são reveladas para que os participantes determinem explicitamente e em comum acordo um conjunto de atributos, quasi-identifiers (QIDs), a serem utilizados na tarefa de REPP [13].

Contudo, estas informações podem ser utilizadas por participantes maliciosos (também chamados de adversários) para quebrar o sigilo e a privacidade dos dados dos demais participantes. Por exemplo, considere a base de dados com informações (anonimizadas) sobre pacientes com o diagnóstico de AIDS ilustrada na Figura 1.1a. Neste exemplo, um atacante pode utilizar a semântica dos dados para reidentificar os dados anonimizados, pois, conhecendo a técnica utilizada para anonimizar os dados e sabendo que, para o campo UF existem apenas 27 valores, que para cada UF existem (em média) 250 cidades e que para cada cidade existem 30 bairros, o atacante pode gerar uma lista de possessíveis valores para cada atributo e reidentificar os atributos. Para reidentificar o nome e o sobrenome dos pacientes o atacante pode utilizar as informações de uma lista telefônica e testar os valores anonimizados

será protegida. Em outras palavras, descreve os deveres que acompanham a divulgação de informações não públicas a terceiros em um relacionamento profissional, legal ou contratual [28].

de todos os residentes de um bairro, resultando na reidentificação de todos os pacientes com diagnóstico de AIDS (Figura 1.1b).

(a) Dados anonimizados

Nome	Sobrenome	Sexo	Bairro	Cidade	UF
0xfa3	0x552	0xfff	0xbbe	0x0d2	0x3ea
0xb4d	0x678	0x111	0xe33	0x31e	0x5c4
0xefa	0xae1	0x111	0xp44	0xc4d	0x3ea

(b) Dados reidentificados

Nome	Sobrenome	Sexo	Bairro	Cidade	UF
Maria	Lima	F	Bessa	João Pessoa	PB
José	Silva	M	Espinheiro	Recife	PE
Abilio	Maia	M	Prata	Campina Grande	PB

Figura 1.1: Base de dados com informações anonimizadas (a) sobre pacientes com o diagnóstico de AIDS que foram reidentificados (b) utilizando um ataque de dicionário.

Além dos problemas de privacidade, problemas como heterogeneidade das bases de dados, as quais, na maioria das vezes, foram projetadas de forma independente, utilizando diferentes modelos de dados (com diferentes representações para os mesmos conceitos do mundo real) e tecnologias (i.e., banco de dados relacional, banco de dados orientados a grafos, entre outras), são tratados de forma manual por analistas de dados na REPP [6, 3, 13, 68].

Desse modo foi identificado como pesquisa científica a proposição de uma abordagem para parear os atributos, ou seja, uma abordagem capaz de identificar os atributos (QIDs) que serão utilizados na tarefa de REPP. Tal abordagem deve: i) considerar a privacidade dos dados, e ii) ser incorporada como uma etapa preliminar ao fluxo da REPP.

1.2 Objetivos

O objetivo geral deste trabalho é propor uma abordagem (semiautomática) capaz de realizar o pareamento de atributos (identificar os QIDs) a serem utilizados durante a tarefa de REPP, preservando a privacidade dos dados. A abordagem deve aumentar a segurança da REPP, eliminando a necessidade de cada participante revelar explicitamente informações sobre os atributos de suas entidades. Desse modo, o objetivo geral do trabalho consiste em avaliar

se é possível identificar QIDs, em bases de dados distintas, preservando a privacidade dos dados em um contexto de REPP.

Considerando o objetivo geral proposto, este trabalho possui os seguintes objetivos específicos:

i propor uma alteração no fluxo tradicional da REPP de modo a adicionar uma etapa preliminar (etapa de Apresentação) na qual serão determinados os atributos QIDs a serem utilizados nas demais etapas da REPP;

ii propor uma abordagem para parear os atributos armazenados em bases de dados distintas que preserve a privacidade dos dados e possa ser utilizada em grandes bases de dados;

iii avaliar a abordagem proposta em termos de eficácia, eficiência e privacidade.

1.3 Escopo

Por apresentar desafios endereçados por diferentes áreas de pesquisa (Integração de dados, Segurança de dados e Privacidade de dados), este trabalho propõe uma abordagem para identificar os atributos similares das entidades em um contexto de REPP com algumas restrições quanto i) ao formato utilizado pelos participantes para representar as entidades (detalhado no Capítulo 4), e ii) quanto ao comportamento dos adversários, chamado de Modelo de Adversário apresentado no Capítulo 2.

Para tal, este trabalho se limita a identificar pares de atributos similares pertencentes a entidades armazenadas em bases de dados distintas, que possam ser utilizados como QIDs na tarefa de REPP. Além disso, este trabalho desconsidera a semântica e o nome dos atributos, utilizando apenas os valores dos atributos das entidades.

1.4 Relevância

O acesso a dados privados, ou seja dados que necessitam que a privacidade e confidencialidade sejam preservados, é regulamentado por leis e normas na União Europeia [22] e nos Estados Unidos da América [66]. No Brasil, o acesso a informações privadas é regulamentado para dados médicos [17], fiscais e financeiros [10]. Contudo, no Brasil, há um Projeto

de Lei (PL 4060/2012) [48] que visa tornar as informações, que identifiquem e descrevam aspectos pessoais, em dados privados, com o objetivo de regulamentar a troca, divulgação e utilização desses dados por entes públicos e privados. A implantação deste PL implicará que empresas, que atualmente utilizam dados de clientes coletados por múltiplas organizações para realizar estudos de mercado e marketing, não poderão mais realizar estes estudos, pois o processo de identificação dos QIDs estaria em desacordo com a lei, reforçando assim a importância deste trabalho para a REPP.

Outro fato relevante é que a utilização de dados privados em diferentes áreas do conhecimento vem sendo descrita em diversos trabalhos científicos, porém, em todos os trabalhos a seleção dos atributos que serão utilizados como QIDs é realizada por analistas de dados e com a divulgação de informações sobre os esquemas e atributos das entidades [14, 38, 70]. Desse modo, a proposição de uma abordagem semiautomática para parear os atributos ressalta a importância deste trabalho por; i) reduzir a quantidade de informações (*e.g.* semântica, tipo do dados, entre outros) divulgadas entre os participantes da REPP, e ii) flexibilizar sua utilização por diversas técnicas de REPP.

1.5 Contribuições

A principal proposta deste trabalho consiste em aumentar a segurança da tarefa de REPP, sem alterar a eficiência e a eficácia, por meio de uma abordagem que identifique os QIDs utilizados durante a tarefa. Para tal, foram desenvolvidos:

i Uma nova etapa (chamada de Etapa de Apresentação) no fluxo tradicional de REPP, que tem a finalidade de identificar e selecionar os QIDs a serem utilizados durante a execução da tarefa de REPP;

ii Uma abordagem a ser empregada na Etapa de Apresentação, a qual elimina a necessidade de os participantes terem que divulgar informações (sobre os atributos que compõem as entidades) que possam ser utilizadas para quebrar a privacidade dos dados originais;

iii Uma estratégia para seleção do número de elementos da amostra (amostragem) que, no contexto da REPP, aumenta a eficiência, eficácia e a segurança da abordagem proposta para a etapa de Apresentação.

Para validar as contribuições foram realizadas avaliações experimentais que investigam a eficácia, privacidade e eficiência da abordagem proposta no trabalho. Tais investigações indicam que é possível aplicar a abordagem proposta para auxiliar a identificação de QIDs a serem utilizados na tarefa de REPP.

1.6 Resultados Alcançados

Em 2016 foi publicado um artigo resumido no 31º Simpósio Brasileiro de Banco de Dados (SBBD) com o título "Avaliação Empírica de Técnicas de Comparação Privada Aplicadas na Resolução de Entidades". Este artigo recebeu menção honrosa como um dos melhores artigos resumidos do simpósio. Em 2017, o artigo "Blind Attribute Pairing for Privacy-Preserving Record Linkage" foi um dos quatro artigos aceitos na trilha "Database Theory, Technology, and Applications - DTTA" apresentado no 33rd ACM/SIGAPP Symposium on Applied Computing (SAC'18).

1.7 Organização do Trabalho

Este documento está dividido da seguinte forma: no Capítulo 2 serão apresentados os tópicos necessários para promover o embasamento teórico para entendimento do trabalho; no Capítulo 3 serão descritos, comentados e comparados os trabalhos relacionados ao trabalho proposto; no Capítulo 4 será descrita a abordagem proposta; no Capítulo 5, será descrita e discutida a avaliação de desempenho; e, por fim, no Capítulo 6 serão apresentadas as conclusões e os trabalhos futuros.

Capítulo 2

Fundamentação Teórica

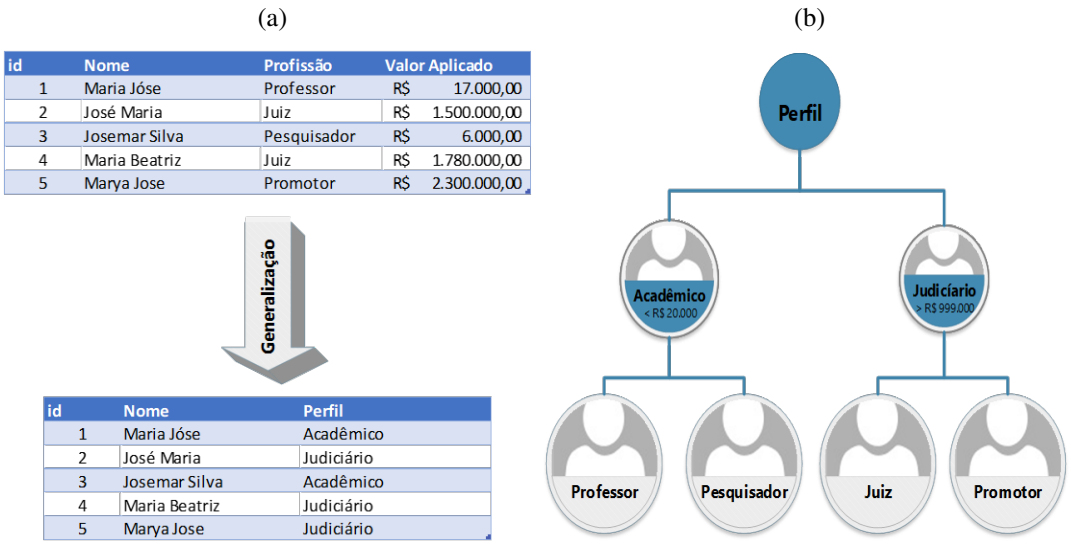
Este capítulo aborda os conceitos básicos necessários para o entendimento do trabalho. Os principais conceitos apresentados são: i) Anonimização de dados, ii) Resolução de Entidades com Preservação de Privacidade, iii) Modelos de Adversários, iv) Protocolos de Resolução de Entidades com Preservação de Privacidade, v) Ataques a dados privados, vi) Avaliação de Privacidade no Contexto da REPP, e vii) Assinatura de dados.

2.1 Anonimização de dados

O objetivo da anonimização de dados é preservar a privacidade e o sigilo da informação. Para tal, as técnicas de anonimização escondem e/ou mascaram o valor original da informação. No contexto da REPP, uma variedade de técnicas vem sendo empregadas, com destaque para duas famílias de técnicas: *Generalização* e *Perturbação de dados* [13, 70].

As técnicas de *Generalização* de dados substituem os valores originais do dado por um valor taxonômico mais amplo, ou seja, valores menos específicos, mas semanticamente consistentes, que os representam. Por exemplo, a Figura 2.1 representa a generalização de uma base de dados sobre investimentos de servidores públicos em uma instituição financeira (Figura 2.1a). O valor aplicado e a profissão foram generalizados de acordo com o Perfil profissional da Figura 2.1b, preservando assim o sigilo fiscal dos indivíduos.

Por sua vez, as técnicas de *Perturbação de dados* (também conhecidas como técnicas



id	Nome	Profissão	Valor Aplicado
1	Maria Jóse	Professor	R$ 17.000,00
2	José Maria	Juiz	R$ 1.500.000,00
3	Josemar Silva	Pesquisador	R$ 6.000,00
4	Maria Beatriz	Juiz	R$ 1.780.000,00
5	Marya Jose	Promotor	R$ 2.300.000,00

id	Nome	Perfil
1	Maria Jóse	Acadêmico
2	José Maria	Judiciário
3	Josemar Silva	Acadêmico
4	Maria Beatriz	Judiciário
5	Marya Jose	Judiciário

Figura 2.1: Exemplo da generalização de uma base de dados (Figura 2.1a) utilizando a árvore taxonômica (Figura 2.1b) a qual foi construída para generalizar o valor investido e a profissão dos indivíduos.

de Mascaramento de dados) consistem em alterar os valores originais dos dados, de modo que: (1) um dado anonimizado não possa ser indevidamente mapeado para o dado original e (2) ao se calcular a similaridade entre dois dados anonimizados, o resultado seja o mesmo ao se calcular a similaridade entre seus respectivos valores originais. Dentre as técnicas de Perturbação de dados utilizadas em REPP, é possível destacar [70]:

i *Random Data Perturbation*: esta técnica adiciona ruído, de forma randômica, aos dados com o objetivo de proteger a privacidade dos dados originais. O ruído pode ser adicionado de duas maneiras distintas: i) nos dados (por exemplo, alternando algumas letras dos atributos de uma entidade), ou ii) na base de dados, com a adição de registros sintéticos (para mitigar ataques de frequência, detalhados na Seção 2.5);

ii *Codificação Fonética*: utiliza a representação fonética (pronúncia) das palavras para mascarar seus valores originais. De acordo com o estudo de Karakasidis [36] esta técnica preserva a privacidade dos dados e apresenta um bom desempenho por utilizar uma estratégia simples de anonimização;

iii *Secure hash enconding*: esta técnica utiliza funções *hash* de mão única[1] (*one-way*

[1]Funções *hash* de mão única, são funções (f(x) = y) que computam, em tempo polinomial, um valor *hash*

hash function) [13] para mapear um dado em um valor *hash* (por exemplo, "João" em "71dd3cdf1ea0"), de modo que, de posse apenas do valor *hash* ("71dd3cdf1ea0"), um adversário não possa inferir qual o valor original correspondente. O *Message Digest 5* (MD5) e *Secure Hash Algorithm* (SHA-1 e SHA-256) são os algoritmos mais conhecidos e utilizados na REPP [13, 70];

iv *Multi-valued Order Preserving Encryption* (MV-OPE): esta técnica utiliza algoritmos de criptografia que possibilita dados (numéricos) iguais tenham suas representações anonimizadas diferentes. Em outras palavras, ao utilizar MV-OPE em dois dados numéricos iguais ($X = Y$) os seus valores anonimizados serão diferentes ($E_k(X) \neq E_k(Y)$) [33]. Tal característica é útil para evitar que ataques de frequência (que serão apresentados na Seção 2.5) sejam executados sobre um conjunto de dados anonimizados. Além da comparação exata de valores (assim com a *Secure hash enconding*) esta técnica mantem a ordem dos valores após a anonimização, ou seja, dados dois valores $X > Y$, a ordem dos seus valores anonimizados pode ser testada ($E_k(X) > E_k(Y)$) [58]. No entanto, o MV-OPE não é amplamente utilizada no contexto da REPP, pois o MV-OPE só permite a comparação exata dos valores anonimizados e não é possível utilizá-lo com dados textuais;

v *Criptografia Homomórfica*: emprega algoritmos de criptografia com chaves simétricas para realizar operações algébricas (como adição privada e produto privado) em dados numéricos encriptados, permitindo assim ordenar e calcular (com precisão) a diferença dos dados anonimizados [55];

vi *Filtros de Bloom*: utiliza funções de mão única para anonimizar dados textuais [9]. Esta técnica permite que dados (anonimizados) possam ser comparados de maneira aproximada, por exemplo a comparação dos valores anonimizados dos nomes "João"e "Joao"retornaria um valor de similaridade de 0,8.

Por possibilitar a comparação aproximada de dados anonimizados e apresentar um baixo custo computacional, o Filtro de Bloom é a técnica de anonimização mais utilizada em

de tamanho fixo (y). Esta função deve ser difícil de inverter, ou seja, dada a imagem da função (valor hash) a probabilidade de encontrar uma pré-imagem (f(x)) deve ser desprezível [13].

REPP [20, 38, 64, 70]. A seguir serão apresentadas em detalhes as principais técnicas de anonimização de dados: o Filtro de Bloom e a criptografia homomórfica.

2.1.1 Filtro de Bloom

Na década de 1990, pesquisadores franceses utilizaram funções *hash* de mão única (SHA1 e MD5) para anonimizar os dados de entradas de estudos epidemiológicos. Tais funções transformam um *String* em um valor *hash*, porém, como ilustrado no Tabela 2.1, a alteração de um único caractere faz com que a função gere um valor *hash* diferente [13]. Ou seja, a utilização de funções *hash* de mão única permite apenas a realização de comparações exatas de dados anonimizados.

Tabela 2.1: Exemplo de utilização de funções *hash* de mão única para anonimizar um String. Como pode ser percebido, a alteração de um caractere minúsculo para um maiúsculo resulta em um valor *hash* distinto.

String	Hash (MD5)
João	09210caa332007b2281ecc88f725d88c
JOÃO	b2478a4a0e4b1dc878f0bd0fb09c72f4
Joao	dccd96c256bc7dd39bae41a405f25e43
JOAO	97ba81591615d92c83d8cf2b6028ef4d

Para possibilitar a comparação aproximada de dados textuais (anonimizados), Schnell [64] propôs a utilização de Filtros de Bloom. O Filtro de Bloom é uma estrutura de dados probabilística que utiliza funções *hash* de mão única para verificar se um determinado elemento pertence a um conjunto (ou não). Desse modo, o Filtro de Bloom é composto por um *array* de bits no qual todos os bits têm o valor zero no momento da criação e, à medida que os elementos são inseridos nele, as funções *hash* de mão única indicam quais posições do *array* devem ser alteradas para representar o elemento recém inserido.

Assim, para que um dado textual seja anonimizado por um Filtro de Bloom, é necessário transformar o dado original em um conjunto de *substrings* (*q-grams*), para que cada *q-gram* seja inserido no Filtro de Bloom. Por exemplo, a Figura 2.2 ilustra a inserção dos nomes

ANA e ANE nos Filtros de Bloom A e B. Primeiramente, os nomes são transformados em *bi-grams* e, em seguida cada bigrama é mapeado por uma função *hash* de mão única para uma posição do filtro, que tem o seu valor alterado para 1.

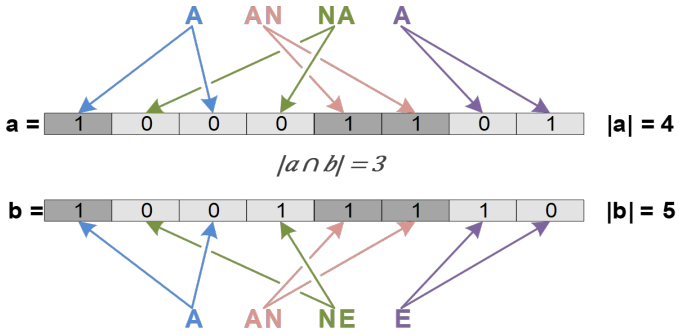


Figura 2.2: Exemplo da inserção dos nomes ANA e ANE em dois Filtros de Bloom.

A similaridade entre dados armazenados em Filtros de Bloom é calculada utilizando funções de distância baseadas em *Token*, como a função de *DICE*, cuja similaridade é dada pela Equação 2.1.

$$DICE = \frac{2 \times |a \cap b|}{|a| + |b|} \tag{2.1}$$

, onde $(|a \cap b|)$ é o número de posições com o valor *1* que coincidem nos dois filtros e $|a|$ e $|b|$ representam o número de total de *1s* em cada filtro.

A Figura 2.2 ilustra a comparação dos nomes ANA e ANE, onde no primeiro momento os nomes são inseridos nos filtros A e B e, em seguida, as posições com o valor 1 são computadas em cada filtro ($|a| = 4$ e $|b| = 5$). Por fim, o número de posições com o valor 1, que coincidem nos dois filtros, é contabilizada ($|a \cap b| = 3$) e aplicado na Equação 2.1, resultando em um valor de similaridade igual a 0,67.

O Filtro de Bloom é a técnica de anonimização mais utilizada em pesquisas sobre REPP [13, 15, 70, 38] por: i) apresentar um baixo custo computacional, ii) impedir que um dado anonimizado seja reidentificado, e iii) possibilitar a comparação aproximada de dados textuais anonimizados.

2.1.2 Criptografia Homomórfica

A Criptografia Homomórfica (CH) [55] é uma técnica de criptografia que utiliza duas chaves, uma para encriptar os dados e outra para a operação oposta. A função possibilita a realização de operações algébricas como adição privada ($\oplus$) e produto privado ($\times$), em dados numéricos encriptados, permitindo assim ordenar e calcular (com precisão) a diferença dos dados. Seja E_k uma função CH que cifra os dados, então E_k possui as seguintes propriedades:

$$\begin{aligned} E_k(x \oplus y) &= E_k(x) \oplus E_k(y) \\ E_k(x) \otimes c &= E_k(x \times c) \end{aligned} \tag{2.2}$$

, onde c é uma constante não encriptada.

A utilização de funções CH permite que a similaridade entre dados não textuais seja calculada utilizando funções especializadas, mantendo a privacidade dos dados e do resultado da comparação [55]. Contudo, de acordo com o trabalho de Nóbrega et. Al. [52] o custo computacional (tempo de execução) de se realizar comparações de dados utilizando CH é, em média, cinco vezes maior do que a utilização de Filtro de Bloom.

2.2 Resolução de Entidades com Preservação de Privacidade

Identificar entidades (e.g., pessoas, publicações, imóveis, entre outros) que se referem a um mesmo objeto do mundo real em múltiplas bases de dados é uma tarefa conhecida por Resolução de Entidades (RE). Esta tarefa tem a sua dificuldade aumentada pela necessidade crescente da identificação de entidades duplicadas em bases de dados com informações privadas e/ou sigilosas, ou seja, informações que não podem ser compartilhadas, a citar, dados médicos, financeiros, entre outros [13, 35, 57].

No contexto de dados privados, a Resolução de Entidades com Preservação de Privacidade (REPP) compreende as abordagens de RE que preservam a privacidade dos dados. Desse modo, a tarefa de REPP deve assegurar que nenhum dado dos participantes seja revelado aos demais (participantes) envolvidos durante a execução da tarefa de REPP. Em outras palavras, nenhum participante deve ser capaz de aprender ou inferir qualquer informação

sobre os dados privados dos demais participantes. Para assegurar que a privacidade do dado seja preservada, a REPP utiliza técnicas para anonimizar os dados (*data anonymization*) de modo que: i) um dado anonimizado não possa ser mapeado para o dado original e ii) ao se calcular a similaridade entre dois dados anonimizados, o resultado seja equivalente ao se calcular a similaridade entre seus respectivos valores originais [70].

Como nem sempre é possível obter um atributo em comum capaz de identificar uma mesma entidade em múltiplas bases de dados (por exemplo, CPF para a entidade Pessoas), a REPP utiliza um conjunto de atributos das entidades (QIDs) para realizar as comparações necessárias e identificar as entidades que possuem alto valor de similaridade com entidades dos demais participantes. Desse modo, é imprescindível que, antes de iniciar a tarefa de REPP, os participantes acordem em três parâmetros:

i **Entidade**: os participantes devem definir o escopo quanto a entidade que será utilizada na REPP. Por exemplo, os participantes devem concordar em fornecer informações sobre pacientes que foram internados em UTI;

ii **Informações sobre QIDs**: atributos a serem utilizados nas etapas da REPP (por exemplo, nome, sobrenome e naturalidade de um paciente), tipo de dados dos atributos (textual, data ou numérico), tamanho médio dos valores dos atributos (por exemplo, o tamanho do atributo CPF é 11 caracteres) e a ordem dos atributos identificadores;

iii **Parâmetros de anonimização**: algoritmos de anonimização, variáveis de inicialização dos algoritmos (por exemplo, número de bits do Filtro de Bloom), chaves de criptografia, entre outros.

Somente após os participantes concordarem em relação aos três parâmetros supracitados, a tarefa de REPP pode ser iniciada. A Figura 2.3 ilustra a tarefa de REPP para as bases de dados de dois participantes (D_a e D_b). As etapas ilustradas na cor azul utilizam dados originais (não anonimizados) enquanto que as etapas mostradas na cor cinza utilizam dados anonimizados.

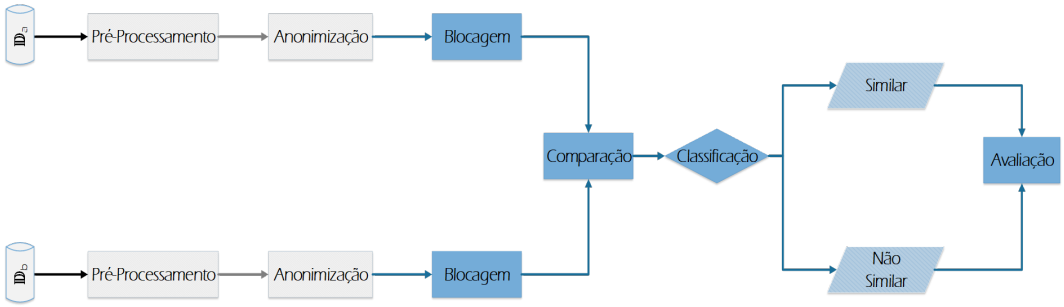


Figura 2.3: Fluxo tradicional da REPP, adaptado de Vatsalan et. al [70].

A etapa de **Pré-processamento** é responsável por corrigir problemas nos dados (como por exemplo, dados faltantes) e resolver problemas de heterogeneidade, representando os dados em um formato que permita a comparação entre entidades [13]. A etapa de **Anonimização** tem a função de mascarar os dados de maneira que possam ser usados pelas etapas subsequentes. Como a anonimização pode ser conduzida por cada participante, de maneira independente, se faz necessário que as organizações envolvidas na REPP troquem informações sobre as entidades a serem comparadas (por exemplo, a ordem dos QIDS), além de parâmetros utilizados pelas funções de anonimização de dados (algoritmos de criptografia, entre outros) [70].

A **Blocagem** (ou Filtragem) tem como objetivo reduzir a quantidade de comparações entre entidades, o que ocorre restringindo-se a execução da REPP para um subconjunto menor de pares de entidades com maior probabilidade de serem consideradas similares, denominados pares candidatos. Como consequência, as etapas subsequentes da REPP passam a consumir menos recursos [45]. As entidades são agrupadas de acordo com algum critério de blocagem como, por exemplo, a primeira letra do nome de um paciente (chave de bloco). As comparações são realizadas apenas entre as entidades que tenham uma chave de bloco em comum, eliminando assim comparações desnecessárias.

Na etapa de **Comparação**, funções de similaridade são aplicadas sobre os dados anonimizados para calcular a similaridade entre os pares candidatos. Tais funções geram valores que representam de forma numérica o grau de similaridade entre as entidades de cada par candidato. Em geral, tais valores são normalizados entre 0 (dissimilar) e 1 (similar) [13]. No contexto de REPP a comparação dos dados das entidades pode ser executada de dois modos. No primeiro modo, os atributos das entidades são concatenados e a comparação é realizada sobre todos os atributos de uma única vez (conhecida por *record comparation*); no segundo

modo, cada atributo é comparado individualmente (*attribute comparation*), portanto a escolha da função de similaridade deve considerar a técnica de anonimização e o tipo do dado (textual, data ou numérico). Ao desconsiderar as características dos dados, a qualidade da REPP pode ser prejudicada. Por exemplo, utilizar uma função para comparação de texto para comparar datas [52].

A etapa de **Classificação** recebe como entrada os valores de similaridade calculados para os pares candidatos da etapa de Comparação e os classifica em similares, dissimilares ou potencialmente similares. Diversos modelos de classificação são utilizados na RE (e.g. Probabilístico, Aprendizagem de Máquina, entre outros). Entretanto, na REPP, o modelo prevalecente é o de limiar (*threshold*), pois os demais modelos, em geral, necessitam dos dados originais (dados não anonimizados) ou de um conjunto de treinamento que normalmente não é possível fornecer no contexto de preservação de privacidade [70]. Ao final da etapa de **Classificação**, uma quantidade limitada de informações é revelada aos participantes da tarefa de REPP: i) o número de entidades que foram classificadas como similares, ii) os identificadores das entidades similares, e/ou ii) um conjunto (pré-selecionado) de atributos das entidades similares [13, 70].

A etapa final da tarefa de REPP é a de **Avaliação**, a qual consiste em examinar o desempenho, a qualidade e a privacidade da REPP. A avaliação da qualidade pode ser aferida utilizando um conjunto auxiliar de dados com gabarito (*gold standard*). É importante salientar que conjuntos de dados privados com gabaritos não são disponibilizados facilmente. O desempenho da REPP pode ser avaliado por meio de experimentos para medir o tempo de execução, consumo de memória, quantidade de comparações, entre outros. Por sua vez, a avaliação da privacidade da REPP por ser realizada por meio de diferentes técnicas (a serem apresentadas na Seção 2.6), como por exemplo o conceito da Probabilidade de Desconfiança (*probability of suspicion*) [69], Entropia [21] ou por meio do Paradigma da Simulação [42]. É importante ressaltar que não há um consenso sobre como avaliar a privacidade de uma tarefa de REPP [70], essa falta consenso torna complexa a comparação das diferentes abordagens propostas para a REPP.

2.3 Modelos de Adversários

No contexto de segurança da informação, adversário (também conhecido por oponente) é um participante malicioso que tenta inviabilizar a utilização de sistemas de criptografia, fazendo com que a privacidade e integridade dos dados protegidas por sistemas de segurança (como, por exemplo, sistemas protegidos por técnicas de criptografia) sejam comprometidas [23]. Assim, as técnicas de REPP consideram três modelos de adversários, para mapear o comportamento dos adversários frente a um protocolo pré-estabelecido entre os participantes do processo [70]:

i **Honesto porém curioso** (HPC): assume que os participantes irão seguir o protocolo corretamente, porém tentarão obter informações adicionais a partir dos dados recebidos durante a execução da REPP. Este modelo não impede que os participantes conspirem uns com os outros (conluio) com o objetivo de descobrir informações sigilosas dos demais participantes [71];

ii **Malicioso**: neste modelo, os participantes se comportam de maneira arbitrária em relação ao protocolo, podendo não o seguir, enviar valores aleatórios, ou até mesmo cancelar o protocolo a qualquer momento. Poucos trabalhos foram realizados em REPP utilizando este modelo de adversário devido a dificuldade de prever como um adversário malicioso pode burlar o protocolo [71];

iii **Dissimulado com auditoria** (*covert and accountable computing*): este modelo surgiu para superar as limitações do HPC, que só pode ser utilizado em cenários onde todos os participantes confiam uns nos outros. O modelo dissimulado com auditoria garante, com alta probabilidade, que os participantes que seguirem o protocolo conseguirão detectar as ações maliciosas dos participantes que não seguirem o protocolo [32, 46, 71].

De acordo com recentes trabalhos de revisão da literatura [13, 70, 71] poucas abordagens consideram os modelos de adversário dissimulado com auditoria e malicioso. O modelo honesto porém curioso continua sendo o modelo de adversário mais utilizado nas abordagens de REPP.

2.4 Protocolos de Resolução de Entidades com Preservação de Privacidade

A REPP tem por objetivo identificar as entidades que se referem ao mesmo objeto do mundo real, em duas ou mais bases de dados, sem que nenhuma informação seja revelada para os participantes durante a execução da REPP. Nesse contexto são utilizados protocolos que podem ser classificados em dois tipos: os que utilizam uma Unidade de Integração (em inglês, "*three-party protocol*") e os que não utilizam Unidade de Integração (também conhecida por "*two-party protocol*") [13].

A Unidade de Integração é uma terceira parte confiável, que tem por objetivo executar as etapas da REPP. Embora a Unidade de Integração possa ser utilizada para executar qualquer etapa da REPP, na maioria das abordagens existentes ela executa as etapas de Comparação e Classificação [70]. A utilização de uma Unidade de Integração em uma tarefa de REPP é ilustrada na Figura 2.4. A REPP é iniciada após a troca inicial de parâmetros (passo 1 da Figura 2.4). Em seguida, os dados são pré-processados e anonimizados pelos participantes para serem enviados à Unidade de Integração (passo 2 da Figura 2.4). Por fim, a Unidade de Integração executa as demais etapas da REPP (Blocagem, Comparação e Classificação) e, ao final, envia aos participantes apenas os identificadores (*id*) das entidades que estão presentes em todas as bases de dados [70].

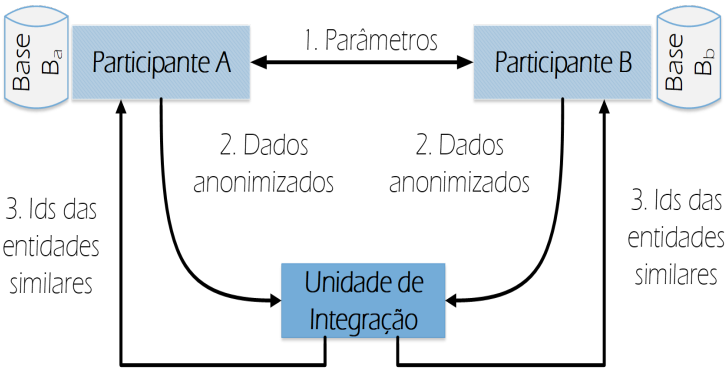


Figura 2.4: Protocolo de REPP que utiliza a Unidade de Integração (three-party protocol).

A Unidade de Integração (terceira parte confiável) deve seguir estritamente o protocolo combinado. Em geral, os protocolos que utilizam a Unidade de Integração são empregados em cenários que considera o modelo de adversário HPC. Como resultado, funções de

anonimização mais simples são utilizadas: i) para reduzir o custo computacional (que se concentra na Unidade de Integração), e ii) pelo fato da Unidade de Integração não poder divulgar as informações que recebe [13].

Por sua vez, os protocolos que não utilizam uma Unidade de Integração (*two-party protocols*) realizam todas as etapas da REPP nos próprios participantes. A Figura 2.5 ilustra, de maneira geral, este protocolo. Como no protocolo anterior, os parâmetros iniciais são trocados e os dados são anonimizados e enviados para os demais participantes (passos 1 e 2 da Figura 2.5). Após a distribuição dos dados (anonimizados) entre os participantes, as demais etapas da REPP são executadas sobre os dados recebidos por cada participante. Ao final, no passo 3, cada participante informa aos demais quais de suas entidades foram classificadas como similares [70].

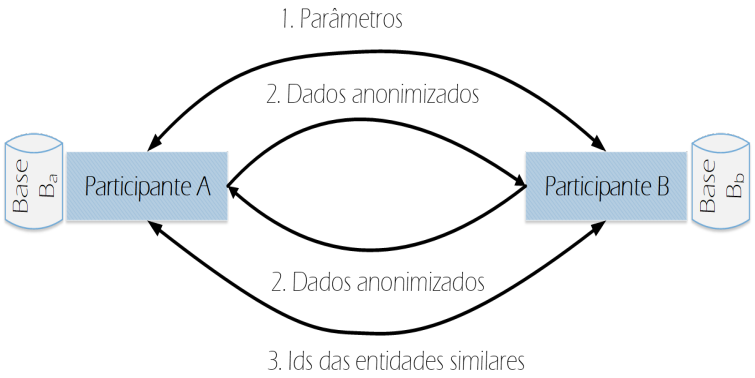


Figura 2.5: Protocolo de REPP que não utiliza a Unidade de Integração (two-party protocol)

Por compartilhar os dados anonimizados com diversos participantes, este protocolo necessita utilizar técnicas de anonimização e comparação privadas mais sofisticadas e, por conseguinte, mais custosas de maneira a assegurar que a privacidade dos dados seja preservada durante a execução da REPP [34]. Dentre as técnicas de anonimização empregadas nesses protocolos é possível citar a Criptografia Homomórfica e técnicas como Secure Multiparty Computation (SMC) [37]. Assim, este protocolo (*two-party protocol*) é mais utilizado em cenários que considera o modelo de adversário Malicioso ou o Dissimulado com auditoria, pois, em geral, por utilizar técnicas de anonimização mais sofisticadas, tendem a ser mais seguros em relação a alguns tipos de ataques de quebra de sigilo e privacidade dos dados [70].

Os protocolos apresentados nesta seção (*two-party protocol* e *three-party protocol*) ainda

podem ser classificados quanto a sua capacidade de serem executados por mais de dois participantes simultaneamente, também conhecido por *multi-party protocols*. A REPP com múltiplos participantes tem por objetivo identificar as entidades similares em múltiplas bases de dados, que pertencem a diferentes participantes, simultaneamente. Como consequência da utilização de multiplos participantes, as etapas de Comparação e Classificação, que antes eram executadas em pares de entidades, necessitam calcular a similaridade de uma entidade em relação a um conjunto de entidades [71].

2.5 Principais Ataques a Dados Privados

Para que a tarefa de REPP seja executada é necessário que dados (anonimizados) sejam compartilhados entre os participantes ou com uma Unidade de Integração. Porém, estes dados anonimizados podem ser utilizados por um adversário para quebrar o sigilo e a privacidade da informação armazenada por este dado. Assim, as abordagens, as técnicas, os protocolos e os dados utilizados na REPP estão vulneráveis aos cinco tipos de ataques descritos a seguir [71]: conluio, frequência, dicionário, composição e criptoanálise.

O ataque de **conluio** ocorre quando dois ou mais participantes se aliam para descobrir informações sobre os dados de outros participantes. Por exemplo, caso a REPP (utilizando um *two-party protocol*) seja executada em um cenário onde os participantes sejam maliciosos, um participante malicioso pode se aliar (conluio) com outro participante com o intuito de violar o sigilo dos dados dos demais participantes [70].

O **ataque de frequência** consiste em observar a frequência em que um determinado valor ocorre em um conjunto de dados anonimizados e compará-la com a frequência em que este valor ocorre em um conjunto de dados conhecido [70]. Por exemplo, em uma base de dados de pacientes diagnosticados com câncer de mama é esperado que o nome Maria apareça com mais frequência que os demais, uma vez que Maria é o nome mais comum para mulheres no Brasil [29]. Desse modo, observando a frequência dos dados anonimizados em uma base de dados, um adversário pode quebrar o sigilo e a privacidade (reidentificar) do dado original.

No **ataque de dicionário**, um adversário anonimiza uma lista (dicionário) de palavras (valores) utilizando diversas técnicas de anonimização até que os valores anonimizados do dicionário coincidam com os valores de registros nos dados atacados [71]. Por exemplo,

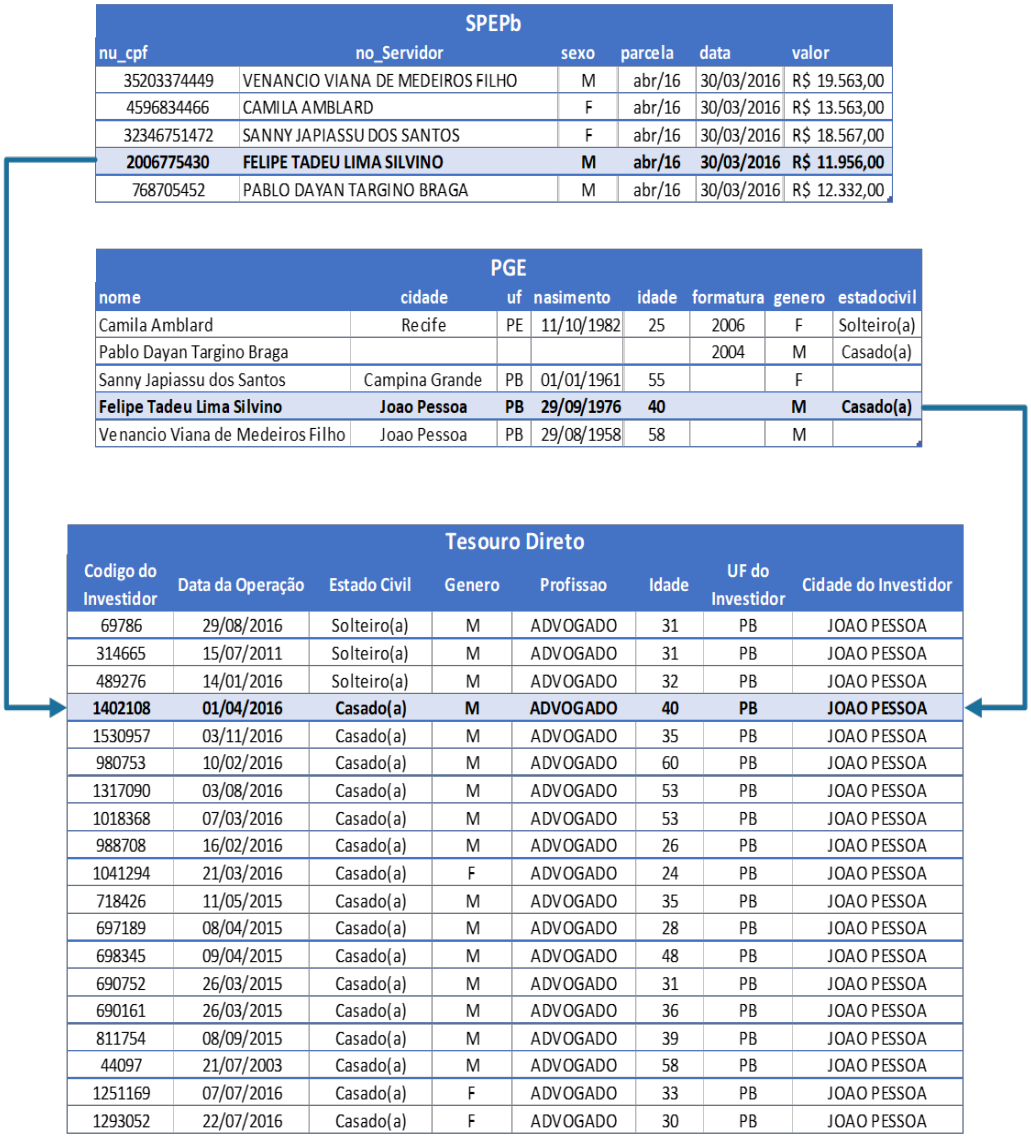
o ataque de dicionário em uma base de dados de pacientes com AIDS pode ser realizado da seguinte maneira: um adversário conhecendo a técnica utilizada para anonimização dos dados, e sabendo que um atributo representa o estado (a sigla) dos pacientes, pode facilmente construir um dicionário, com apenas 23 valores, e identificar o estado (da federação) dos pacientes com AIDS.

No **ataque de composição** informações auxiliares (informações externas) são utilizadas para quebrar o sigilo e a privacidade de dados anonimizados [25]. Parte da fundamentação dos ataques de composição é descrita por Sweeney [65] que, utilizando apenas três atributos (código postal (5 dígitos), gênero e data de nascimento), foi capaz de reidentificar quase 90% da população dos EUA (216 de 248 milhões de indivíduos). Para exemplificar o ataque de composição, a Figura 6 ilustra um ataque a bases de investidores do Tesouro Direto [2] utilizando outra base de dados com os salários dos procuradores do estado da Paraíba (SPEPb)[3] e as informações do site da Procuradoria Geral do Estado da Paraíba (PGE)[4] como fontes de informação externa.

[2]http://dados.gov.br/dataset/operacoes-do-tesouro-direto

[3]https://portal.tce.pb.gov.br/dados-abertos-do-sagres-tcepb/

[4]http://201.18.100.18/portal/conteudos/curriculos/



SPEPb

nu_cpf	no_Servidor	sexo	parcela	data	valor
35203374449	VENANCIO VIANA DE MEDEIROS FILHO	M	abr/16	30/03/2016	R$ 19.563,00
4596834466	CAMILA AMBLARD	F	abr/16	30/03/2016	R$ 13.563,00
32346751472	SANNY JAPIASSU DOS SANTOS	F	abr/16	30/03/2016	R$ 18.567,00
2006775430	**FELIPE TADEU LIMA SILVINO**	**M**	**abr/16**	**30/03/2016**	**R$ 11.956,00**
768705452	PABLO DAYAN TARGINO BRAGA	M	abr/16	30/03/2016	R$ 12.332,00

PGE

nome	cidade	uf	nasimento	idade	formatura	genero	estadocivil
Camila Amblard	Recife	PE	11/10/1982	25	2006	F	Solteiro(a)
Pablo Dayan Targino Braga					2004	M	Casado(a)
Sanny Japiassu dos Santos	Campina Grande	PB	01/01/1961	55		F	
Felipe Tadeu Lima Silvino	**Joao Pessoa**	**PB**	**29/09/1976**	**40**		**M**	**Casado(a)**
Venancio Viana de Medeiros Filho	Joao Pessoa	PB	29/08/1958	58		M	

Tesouro Direto

Codigo do Investidor	Data da Operação	Estado Civil	Genero	Profissao	Idade	UF do Investidor	Cidade do Investidor
69786	29/08/2016	Solteiro(a)	M	ADVOGADO	31	PB	JOAO PESSOA
314665	15/07/2011	Solteiro(a)	M	ADVOGADO	31	PB	JOAO PESSOA
489276	14/01/2016	Solteiro(a)	M	ADVOGADO	32	PB	JOAO PESSOA
1402108	**01/04/2016**	**Casado(a)**	**M**	**ADVOGADO**	**40**	**PB**	**JOAO PESSOA**
1530957	03/11/2016	Casado(a)	M	ADVOGADO	35	PB	JOAO PESSOA
980753	10/02/2016	Casado(a)	M	ADVOGADO	60	PB	JOAO PESSOA
1317090	03/08/2016	Casado(a)	M	ADVOGADO	53	PB	JOAO PESSOA
1018368	07/03/2016	Casado(a)	M	ADVOGADO	53	PB	JOAO PESSOA
988708	16/02/2016	Casado(a)	M	ADVOGADO	26	PB	JOAO PESSOA
1041294	21/03/2016	Casado(a)	F	ADVOGADO	24	PB	JOAO PESSOA
718426	11/05/2015	Casado(a)	M	ADVOGADO	35	PB	JOAO PESSOA
697189	08/04/2015	Casado(a)	M	ADVOGADO	28	PB	JOAO PESSOA
698345	09/04/2015	Casado(a)	M	ADVOGADO	48	PB	JOAO PESSOA
690752	26/03/2015	Casado(a)	M	ADVOGADO	31	PB	JOAO PESSOA
690161	26/03/2015	Casado(a)	M	ADVOGADO	36	PB	JOAO PESSOA
811754	08/09/2015	Casado(a)	M	ADVOGADO	39	PB	JOAO PESSOA
44097	21/07/2003	Casado(a)	M	ADVOGADO	58	PB	JOAO PESSOA
1251169	07/07/2016	Casado(a)	F	ADVOGADO	33	PB	JOAO PESSOA
1293052	22/07/2016	Casado(a)	F	ADVOGADO	30	PB	JOAO PESSOA

Figura 2.6: Exemplo de ataque de composição utilizando informações de duas bases de dados públicas.

No exemplo da Figura 2.6, o ataque de composição é realizado utilizando informações do SPEPb (nome, profissão, data de recebimento do salário) do site da PGE (cidade, estado, idade e estado civil) para reidentificar os investimentos que um indivíduo realizou em títulos do Tesouro Nacional Brasileiro (Tesouro Direto). Assim, combinando as informações do TCE e da PGE, verifica-se que o investidor 1402108 possui a combinação única de seis atributos (estado civil, sexo, idade, estado, cidade e data do investimento) com um servidor da procuradoria do estado. Essa combinação única de informações de múltiplas bases de dados permite que um adversário quebre a privacidade dos dados.

Por fim, os ataques de **criptoanálise** utilizam a combinação de múltiplos ataques para obter informações sobre a chave/algoritmo utilizado para anonimizar os dados. Os Filtros de Bloom podem ser vulneráveis a este tipo de ataque a depender dos parâmetros de anonimização e do volume de dados utilizados. Em razão disso, os participantes trocam informações/parâmetros sobre o comprimento (número de bits) e algoritmos utilizados nos Filtros Bloom [40, 50].

2.6 Avaliação de Privacidade no Contexto da REPP

Ao executar uma tarefa de REPP a privacidade das informações trocadas (ou seja, da entidades) deve ser avaliada com intuito de evitar que estas informações sejam divulgadas inadvertidamente durante a execução da tarefa. Desse modo, algumas técnicas foram propostas para realizar a avaliação da privacidade, a seguir são apresentadas as principais técnicas de avaliação de privacidade (Paradigma da Simulação, Analise de Frequência e *Disclosure Risk*).

O Paradigma da Simulação avalia a privacidade de uma tarefa de REPP com base nas mensagens trocadas durante a execução de uma simulação [42]. Esta técnica considera uma abordagem segura (ou seja, é capaz de preservar a privacidade das informações) se, durante a simulação, um participante não for capaz de inferir qualquer informação sobre os dados (entidades), exceto a sua entrada e saída.

A técnica de Análise de Frequência avalia a probabilidade de reidentificação dos dados (entidades) com base em um ataque de frequência, descrito na Secção 2.5, sobre os dados que são trocadas durante a tarefa de REPP. Por fim, a técnica *Disclosure Risk* [69] calcula a privacidade diferencial com base no conjunto de dados trocados durante a REPP como, por exemplo, o número de atributos e entidades duplicadas. Ambas as técnicas (Análise de Frequência e *Disclosure Risk*) avaliam a privacidade com base nos conjuntos de dados trocados, ou seja, a avaliação é realizada com base em uma instância dos dados utilizados como entrada para a tarefa da REPP.

2.7 Assinaturas de Dados

Um conjunto de dados pode ter suas características representadas na forma de Assinaturas de Dados. O conceito de assinatura pode ser utilizado para classificar dados semanticamente similares em um contexto no qual o tipo e as características dos dados armazenados são desconhecidos (ou escondidas). Por exemplo, em uma base de dados relacional, as assinaturas podem ser utilizadas para identificar atributos similares (e.g. CPF, RG, entre outros) em diferentes bases de dados [1].

Na REPP, a utilização de assinaturas permite que características dos dados sejam compartilhadas sem que o sigilo e a privacidade do dado original sejam comprometidos. No entanto, a quantidade de assinaturas (ou seja, o número de características) que pode ser usada para representar um conjunto de dados muitas vezes depende da finalidade da análise a ser realizada.

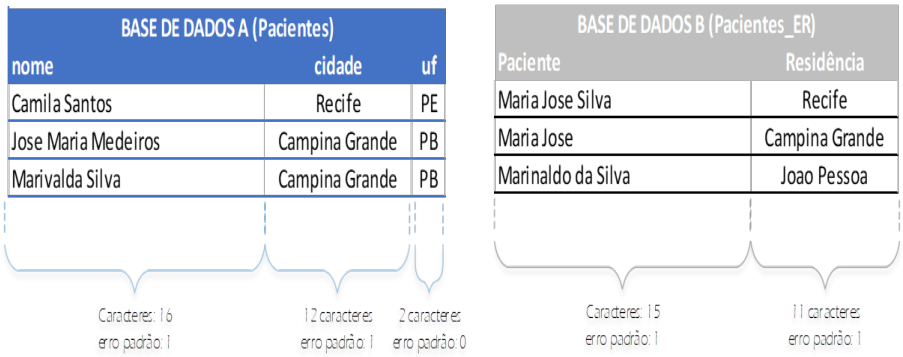


Figura 2.7: Exemplo da utilização de Assinaturas de dados para identificar atributos similares em duas bases de dados.

Para melhor ilustrar a utilidade de Assinatura de Dados no contexto da REPP, considere o exemplo da Figura 2.7. No exemplo, dois hospitais pretendem realizar a REPP em suas bases de dados (base de dados A e B), porém, antes de iniciar a REPP os hospitais devem identificar quais atributos serão utilizados como QIDs. Para tal, foi utilizada uma Assinatura de Dados que representa o número médio e o erro padrão da quantidade de caracteres de cada atributo. Desse modo, ao utilizar a Assinatura de Dados, os hospitais identificaram que os atributos Nome, Paciente e Cidade, Residência fazem referência a mesma informação e devem ser utilizados com QID para a tarefa de REPP.

2.8 Considerações Finais

Neste capítulo foi apresentada uma visão geral sobre os tópicos relevantes para a compreensão desta dissertação. Inicialmente foram apresentados os conceitos relativos à anonimização de dados. Em seguida foram apresentados e discutidos aspectos relativos a REPP, tais como o fluxo tradicional, protocolos, modelos de adversários e principais tipos de ataques a dados. Por fim foi apresentado o conceito de Assinatura de dados, o qual é utilizado nesta dissertação para identificar os QIDs que serão utilizados em uma tarefa de REPP. No capítulo seguinte são apresentados os trabalhos do estado da arte relacionados às contribuições propostas nesta dissertação.

Capítulo 3

Trabalhos Relacionados

Neste capítulo são apresentados e discutidos os principais trabalhos relacionados ao contexto desta dissertação. Pelo fato de propor uma abordagem semiautomática para pareamento de atributos a serem utilizados em uma tarefa de REPP, este trabalho está relacionado a três áreas de pesquisa: i) correspondência entre esquemas, ii) resolução de entidades e iii) protocolos de preservação de privacidade. Assim, a Seção 3 apresenta os trabalhos relacionados dessas áreas, e ao final deste capítulo (Seção 3.4) é realizado um comparativo entre os trabalhos relacionados e o trabalho proposto nesta dissertação.

3.1 Correspondência de Esquemas

Na área de correspondência de esquemas, Bilke e Naumann [8] apresentaram uma abordagem que utiliza as entidades duplicadas presentes em duas bases de dados distintas para realizar (de maneira semiautomática) a identificação de atributos similares. Em outras palavras, a abordagem proposta pelos autores é capaz de identificar atributos que podem ser utilizados como QIDs. A abordagem, denominada de **DUMAS**, não utiliza informações sobre o esquema das entidades, ou seja, o DUMAS desconsidera o nome e o tipo dos atributos que compõem a entidade. Segundo os autores esta solução só é aplicável na presença de dados sobrepostos, ou seja, pelo menos um pequeno número de objetos do mundo real deve estar contido em ambas as bases de dados (registros duplicados).

Para realizar o pareamento de atributos, primeiro o **DUMAS** identifica as entidades duplicadas. Para tal, os valores dos atributos das entidades são utilizados como entrada para um

algoritmo (TF-IDF) para identificar os termos (valores de atributos) mais relevantes de cada entidade. Em seguida, as entidades com maior número de termos em comum são utilizadas para montar uma matriz de similaridade, na qual as colunas representam os valores dos atributos de uma base de dados e as linhas os valores dos atributos da segunda base de dados. Por fim, a similaridade entre os valores das linhas e colunas é computada e utilizada como entrada para o problema de correspondência bipartida ponderada (também conhecido como problema de atribuição) [47], este problema consiste em encontrar as correspondências em um grafo bipartido ponderado. Assim, o pareamento dos atributos (solução ótima) é realizado selecionando-se as colunas e linhas com maior valor de similaridade da matriz.

Por sua vez o trabalho de Jaiswal [30] propõe uma abordagem para identificar as correspondências entre atributos (colunas) de bancos de dados distintos. A abordagem não considera o nome dos atributos (ou seja, não utiliza os nomes dos atributos para realizar o pareamento) para identificar as correspondências entre atributos. Para representar os dados, a abordagem utiliza estratégias diferentes para atributos discretos e contínuos. Por exemplo, para atributos discretos (quantizáveis) a abordagem utiliza o Critério de Informação Bayesiano (que é utilizado para estimar a informação de um dado conjunto de dados) [31], cálculos de densidade dos valores dos atributos e histogramas. Para atributos contínuos (não quantizáveis) a abordagem utiliza Gaussian mixture models (GMM) [31], onde o GMM é um modelo probabilístico para representar os dados das colunas (atributos) com um conjunto de valores em espaço Gausiano. Para identificar as correspondências de atributos, a abordagem propõe a utilização de um algoritmo que utiliza as representações dos atributos (uma para atributos contínuos e outra para discreto) como entrada.

3.2 Resolução de Entidades com Preservação de Privacidade

O trabalho de Scannapieco e Figotin [63], segundo nosso melhor conhecimento, é o único trabalho que aborda o problema de identificar a correspondência entre atributos (QIDs) em um contexto de REPP. O trabalho propõe um protocolo de correspondência entre esquemas que envolve uma terceira parte confiável para fornecer um esquema global que represente as entidades de todos os participantes da REPP. Assim, utilizando o esquema global, os par-

ticipantes devem mapear os atributos de suas entidades (esquemas locais) para o esquema global fornecido pela terceira parte. Esta solução tem a desvantagem que, na prática, não é possível ter um esquema global para todos os domínios de conhecimento. Em outras palavras, por utilizar um esquema global, este protocolo não pode ser utilizado em todos os possíveis cenários da REPP. Além disso, caso seja necessário gerar o esquema global, os participantes terão que fornecer informações sobre seus respectivos esquemas.

Ainda no contexto da resolução de entidade com preservação de privacidade, alguns trabalhos utilizam a criptografia homomórfica para aumentar a privacidade dos dados, com destaque para os trabalhos de Kuzu e Karapiperis [41, 37], descritos a seguir.

O trabalho de Kuzu [41] utiliza uma Unidade de Integração (terceira parte confiável), que tem a função de criar e distribuir as chaves de encriptação (utilizadas pelos participantes para anonimizar os dados), e de realizar a blocagem, comparação e classificação das entidades, enviadas pelos participantes. O trabalho de Kuzu apresenta dois problemas : i) a escalabilidade é comprometida pelo custo computacional da comparação de dados anonimizados com criptografia homomórfica pela terceira parte confiável; e ii) o processo não pode ser realizado por múltiplos (mais de dois) participantes.

Por sua vez, o trabalho de Karapiperis [37] propôs a utilização de *Hamming Locality-Sensitive Hashing* (HLSH) e criptografia homomórfica para integrar dados privados contidos em múltiplas bases de dados. O trabalho propõe que uma terceira parte confiável seja responsável por classificar as entidades e distribuir chaves de encriptação para os participantes anonimizarem os dados. Assim como a etapa de Anonimização, a etapa de Comparação deve ser executada pelos participantes distribuindo o custo computacional dessas etapas. Ainda que o trabalho tenha distribuído as comparações entre os participantes, a escalabilidade da solução proposta é comprometida, pois i) a terceira parte confiável ainda deve realizar a classificação de todos as comparações realizadas pelos participantes e ii) o custo computacional da anonimização e comparação das entidades ainda é elevado em razão da utilização da criptografia homomórfica.

Recentes revisões da literatura [70, 71] apontam como um problema o elevado custo computacional da REPP, em razão da utilização de dados anonimizados em grande volume de dados. Este problema (do elevado custo computacional) foi atacado de modo diferente em diversos trabalhos, principalmente com a proposição de técnicas de Blocagem, para reduzir

o número de comparações da REPP, e utilização de sistemas distribuídos para paralelizar a tarefa de REPP. Alguns desses trabalhos são descritos a seguir.

O trabalho de Franke [24] apresenta uma abordagem que utiliza um *framework* para processamento distribuído de dados (Apache Flink[1]) com o objetivo de viabilizar a execução da tarefa de REPP para grandes bases de dado. A abordagem proposta utiliza um técnica de blocagem baseada em um Locality Sensitive Hashig (LSH) para agrupar as entidades em blocos e distribuir as a execução das comparações para os nós de processamento. Os resultados do experimento demostram que a abordagem proposta foi capaz de distribuir a carga entre os nós de processamento (*Speedup*), de do eficiente e com uma boa qualidade na identificação de entidades duplicadas.

Entre os trabalhos que propõem a utilização de blocagem para dados anonimizados o trabalho de Ranbaduge [60] se destaca por utiliza características do Filtros de Bloom para criar uma árvore, na qual os nós representam blocos de entidades, fazendo com que a etapa de blocagem possa ser conduzida de maneira independente entre os proprietários das bases. Esta técnica possibilita a realização da REPP com múltiplos participantes sem a utilização de uma Unidade de Integração. Para aumentar a qualidade da REPP, o autor propôs uma segunda abordagem distribuída que avalia se os blocos (gerados na etapa de Blocagem) são representativos para o conjunto de dados, podendo até gerar novos blocos para alcançar uma melhor qualidade [61].

3.3 Protocolos de Preservação de Privacidade

Um trabalho de destaque na área de protocolos de preservação de privacidade é o de Qiu et al. [59]. Os autores propõem um protocolo (denominado **SASC**) com três participantes para calcular a interseção de conjuntos de dados privados (similaridade). A similaridade de dois conjuntos de dados é calculada por meio de uma terceira parte confiável que, por sua vez, utiliza criptografia homomórfica [31] para realizar os testes de igualdade e associação dos elementos (entidades) armazenados no conjunto de dados. No entanto, a criptografia homomórfica adotada pelo protocolo SASC possui alto custo computacional e baixa precisão ao comparar dados textuais [52].

[1] http://flink.apache.org

O trabalho de Christen et al. [15] se destaca por investigar a capacidade do Filtro de Bloom em manter a privacidade dos dados anonimizados. O trabalho demostrou que para reidentificar os valores originais dos dados (anonimizados com o Filtro de Bloom) basta que o adversário tenha acesso a uma lista com os possíveis valores que o dado original assume. Para mitigar tais ataques é necessário utilizar implementações do Filtro de Bloom que aumentam a privacidade, reduzindo a qualidade de comparação dos dados, como descritos nos trabalhos de Durham [20] e Alaggan et al. [2], ou utilizar a comparação que concatena todos os atributos de uma entidade em um único registro (comparação baseada em registro). A investigação realizada por Christen et al. reforça a necessidade de ocultar o esquema durante a fase preliminar (troca de informações sobre os QIDs) da REPP, pois um adversário pode utilizar a semântica do atributo para gerar uma lista dos possíveis valores e utilizando o ataque de criptoanálise pode reidenficar facilmente os valores dos atributos.

3.4 Comparativo dos Trabalhos Relacionados

As abordagens propostas por Bilke e Naumann [8] e Jaiswal [30] são capazes de identificar com eficácia atributos similares de diferentes bases de dados, porém não podem ser utilizadas em tarefas de REPP por não considerarem a privacidade dos dados. Já o trabalho de Qiu et al. [59] considera a privacidade dos dados, mas apresenta uma baixa precisão e eficácia quando empregado na REPP [52]. Por sua vez, Scannapieco e Figotin [63] propuseram um protocolo apropriado para a REPP, porém este protocolo necessita que os atributos sejam mapeados para um esquema global. A Tabela 3.1 apresenta uma visão geral das diferenças dos trabalhos relacionados ao trabalho proposto nesta dissertação. Para ilustrar as diferenças foram considerados três aspectos importantes inseridos no contexto desta dissertação, sendo estes:

i **Esquema global:** indica se a abordagem utiliza ou não um esquema global para realizar o pareamento de atributos;

ii **Técnica de identificação**: informa se a abordagem utiliza uma técnica semiautomática ou manual para a identificação de atributos similares;

iii **Modelo de adversário**: informa o modelo de adversário considerado no trabalho.

Como nem todos os trabalhos endereçam todos os aspectos apresentados anteriormente, alguns aspectos foram marcados com um hífen (-), indicando que o trabalho não aborda o aspecto. Como exemplo, os trabalhos de Bilke [8] e Jaiswal [30] não consideram a privacidade dos dados, por consequente o modelo de adversário foi marcado com um hífen.

Tabela 3.1: Tabela comparativa das abordagens.

Abordagem	Esquema global	Técnica de identificação	Modelo de adversário
Bilke e Naumann [8]	N	Semiautomática	-
Jaiswal [30]	N	Semiautomática	-
Qiu et al. [59]	N	-	Malicioso
Scannapieco e Figotin [63]	S	Manual	HPC
Este trabalho	**N**	**Semiautomática**	**HPC**

Como ilustrado na Tabela 3.1, a abordagem proposta nesta dissertação difere dos trabalhos relacionados por: i) identificar os QIDs de um modo semiautomático, ii) considerar a privacidade dos dados e do esquema, e iii) não utilizar um esquema global, ou seja, pode ser utilizada em problemas de qualquer domínio de conhecimento.

3.5 Considerações Finais

Neste capítulo foram apresentados os trabalhos científicos de três áreas distintas que tem relação com a abordagem proposta nesta dissertação. O trabalho de Bilke e Nauman [8] foi apresentado em mais detalhes pois será utilizado como comparativo (*base line*) em alguns experimentos e validação da abordagem proposta por esta dissertação. No capítulo seguinte a abordagem proposta nesta dissertação é apresentada em detalhes.

Capítulo 4

Pareamento de Atributos com Preservação de Privacidade

Este capítulo apresenta em detalhes a abordagem proposta para pareamento de atributos (ou seja, identificação de correspondências entre atributos) a serem utilizados na tarefa de REPP. O detalhamento inclui a definição formal do problema, a metodologia utilizada para representar e comparar os atributos, e o algoritmo utilizado para identificar a correspondência dos atributos. Além disso, propõe-se uma modificação no fluxo tradicional da REPP com a adição de uma nova etapa na qual a abordagem proposta será aplicada. O resultado do pareamento será utilizado nas demais etapas da REPP, como por exemplo, nas etapas de Anonimização e Comparação.

4.1 Adição de uma nova etapa ao fluxo tradicional da REPP

Nesta seção é proposta uma nova etapa a ser introduzida no fluxo tradicional da REPP. Esta nova etapa tem por objetivo realizar o pareamento de atributos que serão utilizados nas etapas de Blocagem, Comparação e Classificação da REPP. O pareamento deve ser realizado de modo a preservar a privacidade dos dados e esquemas das bases de dados envolvidas.

Por se tratar de uma etapa preliminar na qual são definidos os atributos (QIDs) utilizados na tarefa de REPP, a etapa proposta é denominada **etapa de Apresentação** (ou Handshake,

em inglês). A Figura 4.1 ilustra o fluxo modificado da REPP originalmente proposto por Vatsalan [70] (Figura 2.3 do Capítulo 2) com a introdução da etapa de Apresentação.

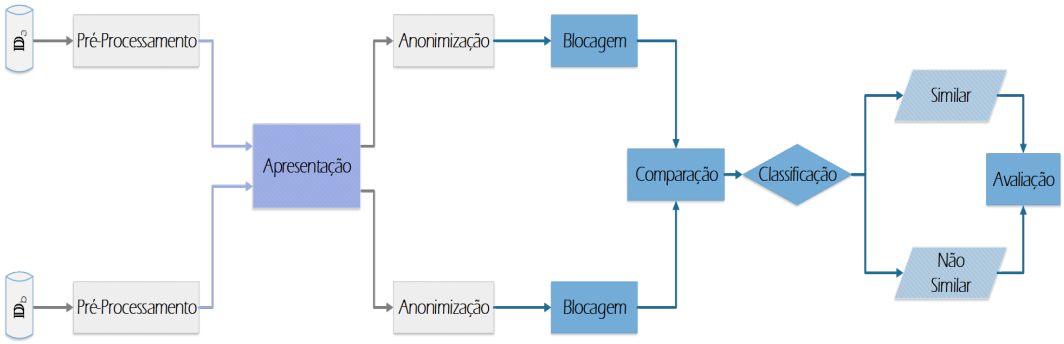


Figura 4.1: Fluxo modificado da REPP com a adição da etapa de Apresentação.

A etapa de Apresentação é executada após a etapa de Pré-processamento, pois esta última etapa (Pré-processamento) processa os dados com o intuito de reduzir problemas de heterogeneidade dos dados, como por exemplo, removendo entidades com atributos faltantes, convertendo dados para um formato padrão, entre outros [13]. Este processamento prévio dos dados (entidades) beneficia o pareamento dos atributos, melhorando a eficácia do pareamento, uma vez que dados com problemas de heterogeneidade tendem a reduzir a eficácia das abordagens [5].

Após a etapa de Pré-Processamento é iniciada a etapa de Apresentação, o resultado desta última etapa (saída) consiste em uma lista com os pares de atributos que devem ser utilizados pelas demais etapas da tarefa de REPP. Assim, em razão da etapa de Apresentação ser responsável por determinar quais atributos serão utilizados nas demais etapas da REPP, a etapa de Apresentação deve seguir algumas diretrizes, sendo elas:

i A etapa de apresentação deve preservar a privacidade dos dados;

ii A etapa de apresentação não deve alterar a funcionalidade das demais etapas da REPP.

Em suma, a adição da etapa de Apresentação deve assegurar que nenhuma informação sobre os dados ou esquema seja divulgada para um adversário, e que as demais etapas da REPP não tenham suas funcionalidades e saídas alteradas. Em outras palavras, a etapa de Apresentação não deve alterar as demais etapas da REPP para assegurar a compatibilidade da etapa de Apresentação com a variedade de técnicas de REPP já propostas [13, 70, 40, 71].

Como principais vantagens da adição da etapa de Apresentação tem-se que os participantes não precisam mais acordar previamente quais atributos serão utilizados como QIDs, pois a etapa de Apresentação identificará e selecionará (preservando o sigilo e a privacidade dos dados e esquemas) quais QIDs devem ser utilizados pela REPP. Outro ponto que deve ser observado é que a etapa de Apresentação não exige alterações nas demais etapas da REPP, ou seja, a etapa de Apresentação pode ser utilizada por qualquer técnica de REPP que siga o modelo descrito na literatura [13, 38, 70, 73]. Nas próximas seções a abordagem proposta (prova de conceito) para a etapa de Apresentação será apresentada e avaliada.

4.2 Definição formal do problema

Antes de iniciar a formalização do problema é importante definir os termos entidade, atributo e base de dados no contexto desta dissertação. Uma entidade representa uma instância de um objeto do mundo real, onde cada característica desse objeto é representada por um atributo. Por sua vez, uma base de dados é composta por um conjunto de entidades. Por exemplo, em bases de dados relacionais, as tabelas representam conjuntos de entidades, os campos das tabelas representam os atributos e as linhas das tabelas (registros) representam as entidades. Ressaltando que a solução descrita nesta dissertação pode ser empregada em bases de dados não relacionais, como por exemplo bases de dados orientados a grafos, documentos, chave valor, entre outras.

No intuito de parear os atributos (armazenados nas diferentes bases de dados), a abordagem assume que as entidades de cada base de dados são representadas por um esquema de tabela única (Figura 4.2), no qual todos os atributos de um conjunto de entidades são armazenados em uma única tabela. Uma tabela única armazena entidades de um domínio comum (por exemplo, hotéis, pacientes ou terroristas) e pode conter diferentes conjuntos de atributos (atributos diferentes) para representar a mesma entidade. Neste contexto, a abordagem proposta deve endereçar os seguintes desafios: i) gerar assinaturas para representar os atributos; e ii) preservar a privacidade do esquema e dados ao executar o pareamento de atributos.

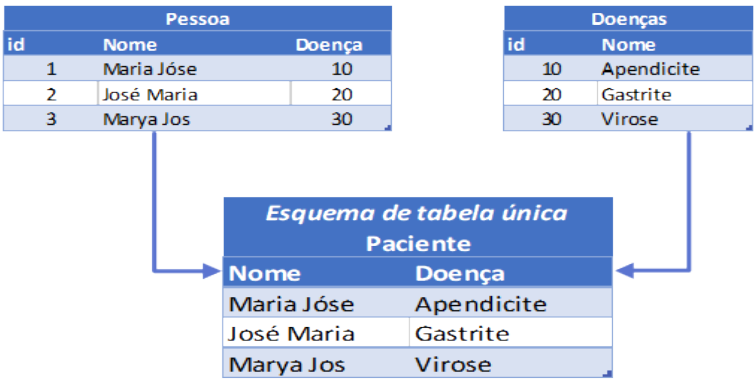


Figura 4.2: Exemplo da conversão de uma base de dados relacional para um esquema de tabela única.

Para formalizar a notação utilizada no restante do trabalho, suponha que cada entidade (e) possui uma lista de atributos $e = [\mathfrak{a}_1, \cdots, \mathfrak{a}_n]$ e é armazenada na base de dados ($\mathbb{D}$), onde $\mathbb{D}$ fornece um esquema de tabela única ($\mathbb{S}$). Assim, dada a base de dados $\mathbb{D}$, é possível recuperar os valores de um atributo usando a função $v(\mathbb{D}, \mathfrak{a}_n)$. Por exemplo, na Figura 4.2 a base de dados dos pacientes ($\mathbb{D}_{pacientes}$) fornece um esquema de tabela única ($\mathbb{S}_{pacientes}$) para a entidade (e) Paciente com os atributos Nome e Doença, então $v(\mathbb{D}_{pacientes}, \mathfrak{a}_{doençа}) = \{Apendicite, Gastrite, Virose\}$.

Com a notação definida, agora os desafios i e ii serão formalizados como problemas. Para realizar o pareamento dos atributos é necessário representar as características de cada atributo como uma assinatura (Ω_n).

PROBLEMA 1 *Dado um atributo* $\mathfrak{a} \in \mathbb{D}$ *como gerar um conjunto de assinaturas* ($\mathfrak{a}^\tau = [\Omega_1, \cdots, \Omega_n]$) *, tal que cada assinatura represente características dos valores do atributo a da base de dados* $\mathbb{D}$, $v(\mathbb{D}, \mathfrak{a}_n)$*?*

As assinaturas ($\mathfrak{a}^\tau$) geradas para cada atributo serão utilizadas para identificar os atributos similares em bases de dados distintas. Por exemplo, na Figura 4.2 as assinaturas geradas para os atributos *Nome* e *Doença* da entidade *Paciente* serão utilizadas para identificar, na base de dados de outro participante da tarefa de REPP, quais atributos são similares e devem ser utilizados como *QIDs* pela tarefa de REPP.

O cojunto de assinaturas é definido como $\mathbb{D}^\tau$, representando os valores dos atributos da base de dados ($\mathbb{D}$), onde $\mathbb{D}^\tau = assinaturas(\mathfrak{a}) | \forall \mathfrak{a} \in \mathbb{D}$. Supondo que $\mathfrak{a}^\tau \in \mathbb{D}^\tau_r$ e $\mathfrak{b}^\tau \in \mathbb{D}^\tau_t$, a comparação entre $\mathfrak{a}^\tau$ e $\mathfrak{b}^\tau$ é realizada usando uma função de similaridade $\psi(\mathfrak{a}^\tau, \mathfrak{b}^\tau)$. Esta

função de similaridade (ψ) retorna um valor entre 0 e 1, onde 1 representa a similaridade máxima entre os atributos comparados e 0, completa dissimilaridade. Assim, para executar o pareamento de atributos mantendo a privacidade do esquema e os dados originais o segundo problema é formalizado como:

PROBLEMA 2 *Dados dois esquemas de tabela única distintos ($\mathbb{S}_r$ e $\mathbb{S}_t$) fornecidos pelas bases de dados ($\mathbb{D}_r^\tau$ e $\mathbb{D}_t^\tau$), um limite de similaridade σ e uma função de similaridade ψ, como identificar correspondências ($\Gamma_{\mathbb{S}_r,\mathbb{S}_t}$) entre atributos de $\mathbb{D}_r^\tau$ e $\mathbb{D}_t^\tau$, tal que* $\Gamma_{\mathbb{S}_r,\mathbb{S}_t} = \langle \mathfrak{a}^\tau, \mathfrak{b}^\tau \rangle \in \mathbb{D}_r^\tau \times \mathbb{D}_t^\tau, \quad \exists \psi(\mathfrak{a}^\tau, \mathfrak{b}^\tau) \geq \sigma$*?*

As soluções para os Problemas 1 e 2 devem considerar três fatores importantes: i) eficiência da abordagem, ii) eficácia do pareamento dos atributos e iii) a privacidade dos dados e esquemas das bases de dados (*riscoPrivacidade*). Desta forma, a abordagem deve considerar um modelo de adversário honesto porém curioso, utilizar um protocolo *three-party protocol* e seguir as seguintes diretrizes:

Dado:	$\mathbb{D}_r, \mathbb{D}_t, \sigma, \psi$
Crie:	$\mathbb{D}_r^\tau, \mathbb{D}_t^\tau$
Encontre:	$\Gamma_{\mathbb{S}_r,\mathbb{S}_t}$
Minimize:	*riscoPrivacidade*$(\mathbb{D}_r^\tau, \mathbb{D}_s^\tau, \sigma, \psi)$
Maximize:	*eficiência*$(\Gamma_{\mathbb{S}_r,\mathbb{S}_t})$
Maximize:	*eficácia*$(\Gamma_{\mathbb{S}_r,\mathbb{S}_t})$

4.3 Abordagem de Referência para o Pareamento Privado de Atributos

Nesta seção será apresentada a abordagem de Pareamento de Atributo às Cegas (PAC), a abordagem semiautomática para o pareamento dos atributos formalizada na seção anterior. A abordagem PAC é dividida em três fases: i) Padronização e Amostragem, ii) Criação de Assinaturas e iii) Pareamento Privado de Atributos. A Figura 4.3 ilustra a execução do PAC para dois participantes (A e B). As fases de Padronização e Amostragem e Criação das

Assinaturas são executadas de forma independente por cada participante enquanto que a fase de Pareamento Privado de Atributos é executada por uma terceira parte confiável (TPC).

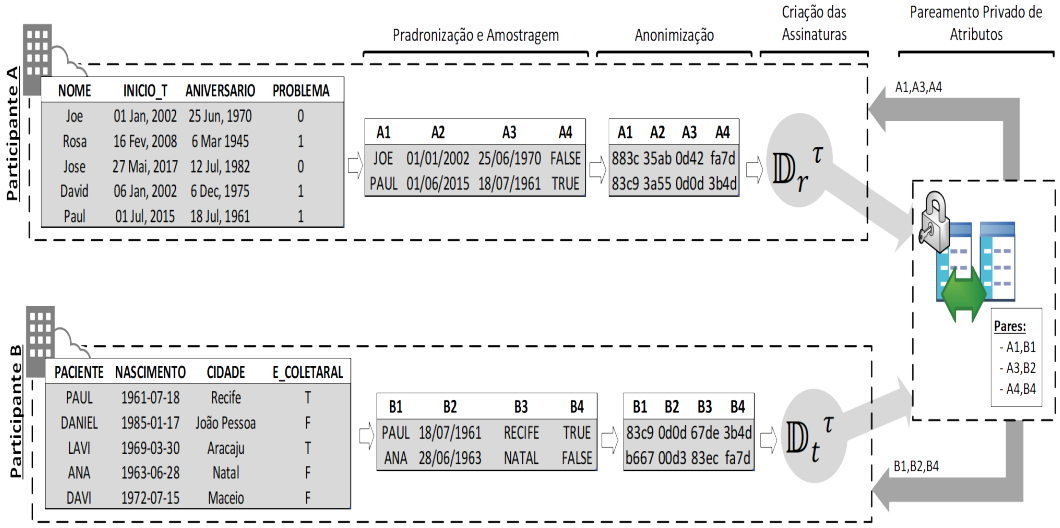


Figura 4.3: Visão geral do PAC.

Inicialmente, cada participante extrai uma amostra das entidades presentes em suas bases de dados, realiza uma padronização dos dados e anonimiza os valores dos atributos. A amostragem tem a função de maximizar a eficiência e a eficácia do PAC (discutido na seção 4.3.1). A Padronização converte os valores dos atributos da amostra para um formato comum, previamente acordado entre os participantes. Após a Padronização, com os dados anonimizados (utilizando o Filtro de Bloom), os participantes geram assinaturas para cada atributo do seu esquema de tabela única, que em seguida são enviadas (as assinaturas) para uma terceira parte confiável que executará a fase de Pareamento Privado de Atributos. Ao final do PAC, a terceira parte confiável envia para os participantes quais dos seus atributos serão utilizados na continuação da tarefa de REPP.

Contudo, antes da execução do PAC, os participantes devem acordar em alguns parâmetros: formato comum dos dados e tamanho do Filtro de Bloom. O formato comum dos dados tem o objetivo de reduzir problemas de relacionados a heterogeneidade dos dados [13]. Por exemplo, na Figura 4.3 os participantes A e B representam de forma distinta as datas para que as datas sejam comparadas adequado os participantes acordam em um formato comum para as datas (ilustrada na etapa de Pré-processamento da Figura). Por sua vez, o tamanho do Filtro de Bloom tem um papel fundamental para a abordagem pois o PAC utiliza o mesmo tamanho de filtro para todos os atributos. A utilização do mesmo tamanho de filtro

previne que adversários utilizem o tamanho do filtro para reidentificar atributos [15].

Assim, mesmo com adoção de um formato comum dos dados, a utilização da abordagem não é capaz de solucionar todo o tipo de correspondência entre atributos, por exemplo, um atributo de uma base (nome completo) pode ter mais de uma correspondência em outra base de dados (nome e sobrenome). Desse modo, o PAC considera apenas correspondências simples de atributos (1:1) na fase de Pareamento Privado de Atributos. Em outras palavras, dados dois esquemas $\mathbb{S}_r \in \mathbb{D}_r$ e $\mathbb{S}_t \in \mathbb{D}_t$, o PAC realiza o pareamento de um atributo do primeiro esquema ($\mathfrak{a} \in \mathbb{D}_e$) a um atributo do segundo esquema ($\mathfrak{b} \in \mathbb{D}_t$).

Nas subseções seguintes são detalhadas a Amostragem, criação e a função de similaridade utilizada para as Assinaturas e por fim é apresentado Pareamento Privado de Atributos, e como a comparação dos atributos é realizada pela terceira parte confiável.

4.3.1 Amostragem

Nesta seção são apresentadas duas estratégias para definição do número de elementos (entidades) das amostras utilizadas pelo PAC. As estratégias aqui apresentadas não propõem uma metodologia para seleção dos elementos das amostras (a citar, aleatória, estratificada ou sistemática) pois a proposição e desenvolvimento de metodologia para selecionar amostras representativas em bases de dados com dados privados é uma tarefa complexa que requer um estudo especifico para conciliar a privacidade dos dados com a representatividade da amostra [7, 60]. Assim, as estratégias aqui propostas para selecionar o número de elementos das amostras podem ser empregadas e adaptadas para serem utilizadas em qualquer metodologia de amostragem. Na avaliação do PAC foi utilizada a metodologia aleatória para selecionar os elementos das amostras.

As estratégias propostas nesta seção têm por objetivo: i) ocultar o número total de entidades presentes nas bases de dados dos participantes (adversários), ii) reduzir o número de mensagens trocadas entre os participantes, e iii) melhorar a eficiência e eficácia do PAC. Ao ocultar o número total de entidades presentes nas bases de dados e reduzir o número de mensagens trocadas pelos participantes, as estratégias reduzem a quantidade de informação disponível para os adversários, dificultando que estes (adversários) utilizem essa informação para inferir algo sobre os valores dos atributos (representados por Assinaturas de Dados) dos demais participantes.

A seguir serão apresentadas duas estratégias (Independente e Conjunta) para selecionar o número de elementos da amostra que utilizam a Terceira Parte Confiável (TPC) para definir o número de elementos das amostras.

Para definir o número de elementos das amostras, a primeira estratégia (**estratégia Independente**) aplica um percentual (ϑ), definido pela TPC, em cada base de dados. Nesta estratégia os participantes enviam o número de entidades presentes em suas bases de dados para a TPC que, utilizando a Equação 4.1, define o número de elementos das amostras para cada base de dados.

$$Amostra = N_{base} \times \vartheta \tag{4.1}$$

Na Equação 4.1, N_{base} representa o número de entidades presentes na base de dados e ϑ o percentual definido pela TPC, ou seja, cada participante extrai uma amostra com o mesmo percentual (ϑ) de sua base, que na pratica faz com que os participantes utilizem amostras com o número de elementos distintos. A utilização da **estratégia Independente** impede que adversários infiram informação sobre o número total de entidades presentes na base de dados dos demais participantes, pois a única informação que o adversário recebe é o número de entidades que devem ser extraídas da sua base de dados (número de elementos da amostra).

Por sua vez, a **estratégia Conjunta** tem o objetivo de utilizar amostras com o mesmo número de elementos para todos os participantes. Para tal, os participantes enviam o número total de elementos de suas bases de dados para TPC que utiliza a Equação 4.2 para definir o número de elementos da amostra.

$$Amostra = N_a + N_b \times \vartheta + err \tag{4.2}$$

Na Equação 4.2, N_a e N_b representam o número de elementos das bases A e B, ϑ representa o percentual escolhido pela TPC, e a variável err representa um erro (aleatório) definido pela TPC para evitar que um adversário possa inferir algo sobre o número de entidades nas bases dos demais participantes.

As mensagens trocadas pelas estratégias Independente e Conjunta são similares tendo como única diferença a equação aplicada (Equação 4.1 ou 4.2) para definir o número de elementos da amostra. As mensagens trocadas entre a TPC e os participantes são ilustradas da Figura 4.4.

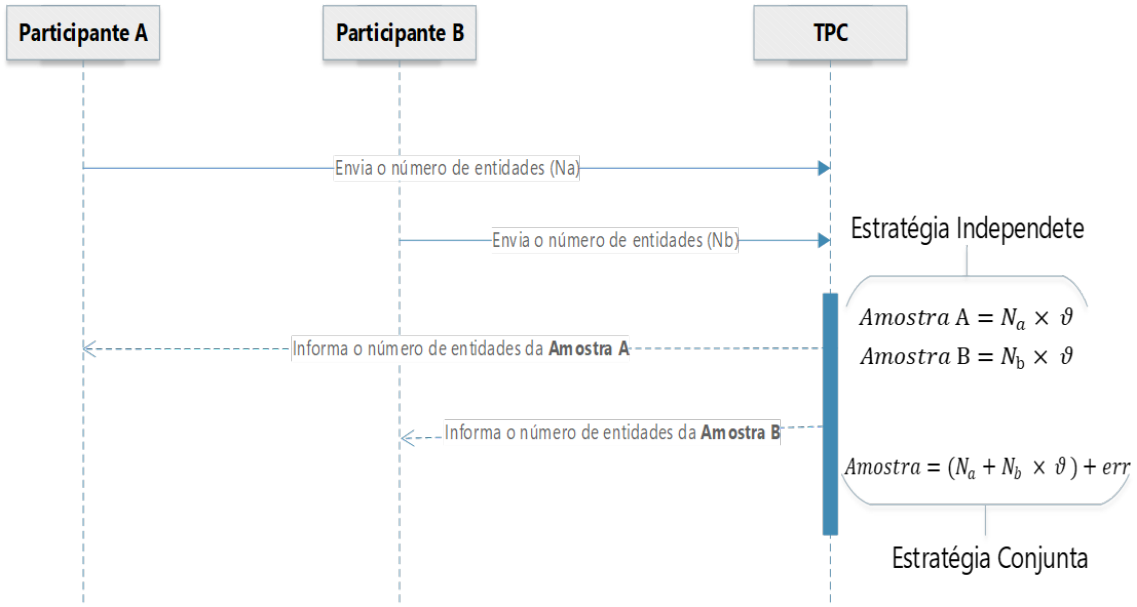


Figura 4.4: Mensagens trocadas durante a execução da estratégia independente e conjunta para escolha do número de elementos da amostra.

Em termos práticos a utilização da **estratégia Independente** faz com que as amostras tenham um número de elementos diferentes. Por sua vez, a **estratégia Conjunta** faz com que as amostras tenham o mesmo número de elementos. A utilização dessas estratégias é investigada no Capítulo 5, onde experimentos foram realizados para avaliar a utilização das estratégias.

4.3.2 Anonimização dos Dados

Nesta seção é realizada uma discussão sobre o processo de anonimização utilizada pelo PAC. Conforme as diretrizes apresentadas na Seção 4.1, a etapa de Apresentação deve ser compatível com a maioria das abordagens de REPP e garantir a privacidade dos dados e esquemas durante o pareamento dos atributos. Assim, para tornar a abordagem proposta compatível com as diretrizes supracitadas, o PAC utiliza um processo de anonimização dos dados diferente do processo de anonimização utilizado na etapa de Anonimização de dados da REPP, pois:

i A utilização do mesmo processo de anonimização da etapa de Anonimização pode inviabilizar o uso do PAC em técnicas de REPP que concatenam todos os atributos (*record comparation*);

ii Redução do custo computacional da etapa de Anonimização da REPP ao evitar que atributos que não pareados sejam anonimizados, comparados e classificados.

iii A utilização de técnicas de anonimização distintas (no PAC e etapa de Anonimização) pode aumentar a privacidade dos dados.

O primeiro item visa assegurar que o PAC seja utilizado pelo maior número possível de técnicas de REPP, pois técnicas que utilizem comparação das entidades inteiras (*record comparation*) pode inviabilizar a utilização do PAC. Como descrito no Capítulo 2, estas técnicas anonimizam as entidades inteiras, concatenam os valores de todos os atributos de uma entidade com uma única linha textual. Por exemplo, a anonimização (utilizando a *record comparation*) da entidade *"Maria Jóse"* da Figura 4.2 consiste em concatenar os atributos *Nome* e *Doença* em um único campo ("Maria JóseApendicite") e anonimizar este campo com um Filtro de Bloom. Desse modo, o PAC não seria capaz de realizar o pareamento de atributo pois todos as entidades seriam representadas com um único atributo, reforçando a necessidade de utilizar um processo de anonimização diferente ao utilizado na etapa de Anonimização.

O item ii considera que, em um processo de pareamento de atributos, é esperado que nem todos os atributos presentes nas bases de dados sejam utilizados para realizar a REPP. Para melhor ilustrar considere o exemplo da Figura 4.3 onde apenas três atributos de cada base (nome, cidade e data de nascimento) serão utilizados na tarefa de REPP, ou seja, a etapa de anonimização não necessita anonimizar todos os atributos das bases, reduzindo assim custo computacional (aumentando a eficiência) da etapa de Anonimização dos dados. Esta redução do custo computacional é especialmente importante ao se trabalhar com grandes bases de dados, pois a anonimização de um único atributo em uma base de dados com milhões de entidades representa uma redução importante do custo computacional [71].

Por fim, a utilização de processos de anonimização distintos pode aumentar a privacidade dos dados (item iii). Para melhor ilustrar esta afirmação, suponha que a REPP seja executada por participantes (adversários) maliciosos, ou seja, onde os participantes não seguiram o protocolo combinado. Assim caso os participantes entrem em conluio e compartilhem informações sobre as assinaturas geradas durante a execução do PAC, estas informações terão pouco utilidade nas demais etapas da REPP, pois os dados foram anonimizados de

forma distinta no PAC e na etapa de Anonimização. Em outras palavras, um adversário terá mais dificuldade em utilizar as informações da PAC para reidentificar as entidades (anonimizadas) durante as demais etapas da REPP.

4.3.3 Criação de Assinaturas

Com os dados padronizados, amostrados e anonimizados, a fase de Criação das Assinaturas é iniciada. Como introduzido no Capítulo 2, as assinaturas representam características dos atributos das entidades armazenadas nas bases de dados, preservando a privacidade dos dados e do esquema. A abordagem proposta considera duas assinaturas para representar os atributos armazenados por uma base de dados, sendo elas:

i **Assinatura de Equivalência da Informação** (AEI): mede a quantidade de informação que cada atributo representa. Considerando o exemplo da Figura 4.3, o atributo nome armazena mais informação (com o maior número de caracteres e com mais valores distintos) do que o atributo problema, o qual apresenta apenas dois valores possíveis "0" ou "1" (com apenas um caractere);

ii **Assinatura de Similaridade dos Dados** (ASD): reduz a dimensionalidade dos atributos de uma base de dados. Representa os valores assumidos por um atributo como um único valor *hash*, $\Omega_{\mathfrak{a}} = gerar\_asd(v(\mathbb{D}, \mathfrak{a}_n))$. Esta assinatura permite calcular a similaridade do conjunto dos valores assumidos por um atributo. Em outras palavras, esta assinatura permite calcular a similaridade dos atributos baseado na intercessão dos seus valores.

A abordagem proposta nesta dissertação utiliza um par de assinaturas (AEI e ASD), $\Omega_{\mathfrak{a}} = [AEI(\mathfrak{a}^{\tau}), ASD(\mathfrak{a}^{\tau})] | \forall \mathfrak{a}^{\tau} \in \mathbb{D}^{\tau}$, para representa os atributo ($\mathfrak{a}$) de uma base de dados ($\mathbb{D}$). A seguir são detalhados os processos de criação e utilização das assinaturas (AEI e ASD).

Assinatura de Equivalência da Informação

Essa assinatura é baseada em um conceito bem conhecido na Teoria da Informação, entropia [18]. A entropia representa a distribuição de informação de uma variável aleatória em um universo que, no contexto desta dissertação, é um atributo. Assim, sendo X uma variável

aleatória discreta com o alfabeto χ e a função de massa de probabilidade $p(x) = Pr\{X = x\}, x \in \chi$, é possível definir entropia, $H(X)$ como na Equação 4.3:

$$H(X) = -\sum_{x \in \chi} p(x) \log p(x) \tag{4.3}$$

Uma vez que a entropia pode ser utilizada para medir a informação de qualquer fonte de informação [18], no contexto desse trabalho, é possível definir que χ (o alfabeto) representa todas as combinações de valores que o Filtro de Bloom pode assumir (número de bits com o valor "1") para cada valor do atributo ($v(\mathbb{D}\tau, \mathfrak{a}^\tau)$. Por sua vez, x representa o número bits com o valor "1" de cada atributo presente na amostra e $p(x)$ é a probabilidade deste atributo (anonimizado Filtro de Bloom) apresentar x *bits* com o valor "1" no alfabeto (χ), ou seja, na amostra. De acordo com Schnell [64] o número de *q-grams* é proporcional ao número de bits com o valor "1". Desse modo, a utilização da entropia no PAC permite avaliar se os valores assumidos por um atributo apresentam uma quantidade de q-grams equivalentes, sem a divulgação de estatísticas (média ou mediana) do número de *q-grams* dos atributos.

Para calcular a similaridade das Assinaturas de Equivalência de Informação, assuma que $\mathfrak{a}^\tau \in \mathbb{D}_r^\tau$ e $\mathfrak{b}^\tau \in \mathbb{D}_t^\tau$, e que a similaridade entre $\mathfrak{a}^\tau$ e $\mathfrak{b}^\tau$ pode ser calculada pela Equação 4.4:

$$H\_sim(\mathfrak{a}, \mathfrak{b}) = \frac{min(H(\mathfrak{a}), H(\mathfrak{b}))}{max(H(\mathfrak{a}), H(\mathfrak{b}))} \tag{4.4}$$

A Equação 4.4 utiliza o valor mínimo da entropia no quociente e o valor máximo de entropia como divisor para assegurar que a função de similaridade (*H_sim*) retorne valores entre 0 e 1, onde 0 representa nenhuma similaridade e 1 similaridade máxima.

Como ilustrado na Equação 4.3, $H(X)$ recebe como entrada o conjunto de valores anonimizados que um atributo assume na base de dados ($\chi = v(\mathbb{D}\tau, \mathfrak{a}^\tau)$). Este conjunto de valores (χ) é utilizado como índice do somatório ($x \in \chi$). Desse modo, no contexto do PAC, é possível afirmar que o número de elementos do conjunto de valores (χ) pode interferir no valor da similaridade, pois como o $H(X)$ utiliza um somatório para iterar sobre estes valores, é possível que conjuntos de dados muito similares apresentem valores de *H_sim* baixo.

A Figura 4.5 ilustrar a utilização da AEI, onde o processo de criação da assinatura é detalhado para dois dos atributos do exemplo da Figura 4.3. Na Figura 4.5 os valores anonimizados dos atributos nome (A1 e B1) e data de nascimento (A3 e B2) são representados

por Filtros de Bloom (de 4 *bits*).

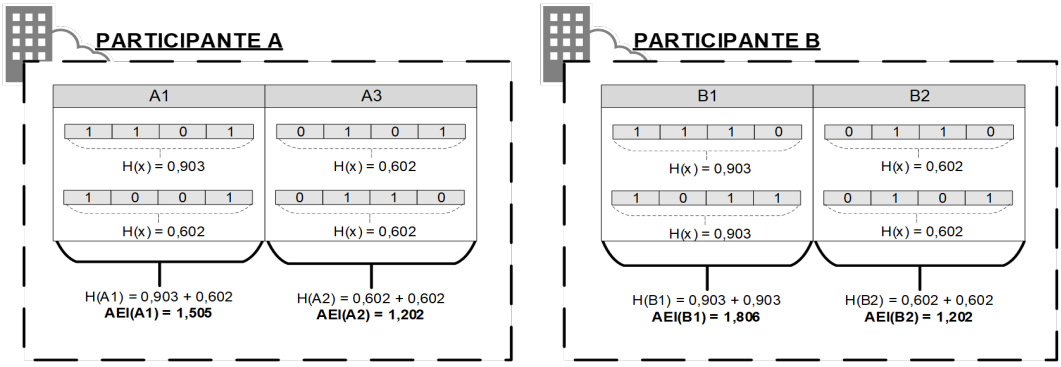


Figura 4.5: Criação da AEI

A assinatura de equivalência de informação é gerada utilizando o valor de entropia dos atributos (Filtro de Bloom), $v(\mathbb{D}\tau, \mathfrak{a}^{\tau})$. Ao aplicar a função de similaridade (*H_Sim*, Equação 4.3), nos valores de AEI, obtém-se o resultado exposto na Tabela 4.1. Nesta é possível perceber que os atributos que representam os nomes (A1 e B1) e data de nascimento (A3 e B2) obtiveram os maiores valores de similaridade das linhas e colunas da Tabela 4.1.

Tabela 4.1: Cálculo de similaridade (*H_Sim*) dos atributos do exemplo.

	B1	B2
A1	**(1,5)/(1,8)=0,83**	1,2/1,5=0,80
A3	1,2/1,8=0,66	**(1,2)/(1,2)=1**

Assinatura de Similaridade dos Dados

Esta assinatura utiliza o algoritmo *MinHash*, proposto por Broder [11], com a finalidade de estimar a similaridade de *Jaccard* (J_{mh}) entre atributos de bases de dados distintas. A ideia básica do *MinHash* é substituir o conjunto dos valores originais por uma representação única, ou seja, a ASD utilizará um único valor *hash* para representar todos os valores assumidos por um atributo em uma base de dados. Para gerar a ASD (Ω_{asd}) do conjunto de valores de um atributo ($\mathfrak{a} \in \mathbb{D}$), o *MinHash* aplica k funções hash (MH_k) em todos os valores que o atributo assume $v(\mathbb{D}\tau, \mathfrak{a}^{\tau})$. Por fim, o algoritmo escolhe a representação mínima destes valores. A Assinatura de Similaridade dos Dados é definida de acordo com a Equação 4.5:

$$\Omega_{asd} = MH_k(val) \forall val \in v(\mathbb{D}\tau, \mathfrak{a}^\tau) \tag{4.5}$$

Para calcular a similaridade entre as ASD é utilizada a medida de similaridade de *Jaccard*. Para tal, admita que os atributos $\mathfrak{a}^\tau \in \mathbb{D}_r^\tau$ e $\mathfrak{b}^\tau \in \mathbb{D}_t^\tau$, utilizando $\mathfrak{a}^\tau$ e $\mathfrak{b}^\tau$ a similaridade de *Jaccard* é expressa pela Equação 4.6 [11]:

$$J_{mh}(\mathfrak{a}^\tau, \mathfrak{b}^\tau) = \frac{MH_k(\mathfrak{a}^\tau) \cap MH_k(\mathfrak{b}^\tau)}{k} \tag{4.6}$$

com um erro de $\theta_j = O(\frac{1}{\sqrt{k}})$

Outro aspecto importante da utilização do algoritmo *MinHash* para calcular a similaridade entre atributos é que, de acordo com o trabalho de Yan et al. [72], o *MinHash* assegura a preservação da privacidade ao utilizar dados anonimizados (e/ou com ruídos) como entrada ou utilizando o algoritmo *Private MinHash Value Generation* (*PrivMinHash*) [72]. Como mostrado nas Equações 4.5 e 4.6, este trabalho utiliza os valores de atributos anonimizados ($\mathfrak{a}^\tau$ e $\mathfrak{b}^\tau$) como entrada para gerar as Assinaturas de Similaridade dos Dados.

A Figura 4.6 ilustra o processo de criação da ASD para os atributos nome (A1 e B1) e data de nascimento (A3 e B2) do exemplo da Seção 4.3 (Figura 4.3). Os valores anonimizados dos atributos são representados por Filtros de Bloom com 4 bits. Para fins deste exemplo, assuma que o algoritmo de *MinHash* utiliza um array com 4 bits (onde inicialmente todos os *bits* assumem valor '0'), e que é utilizado uma única função hash (k) que altera os bits do *array* que representa o *MinHash* para '1' quando todos os bits dos Filtros de Bloom apresentam o valor '1' na mesma posição.

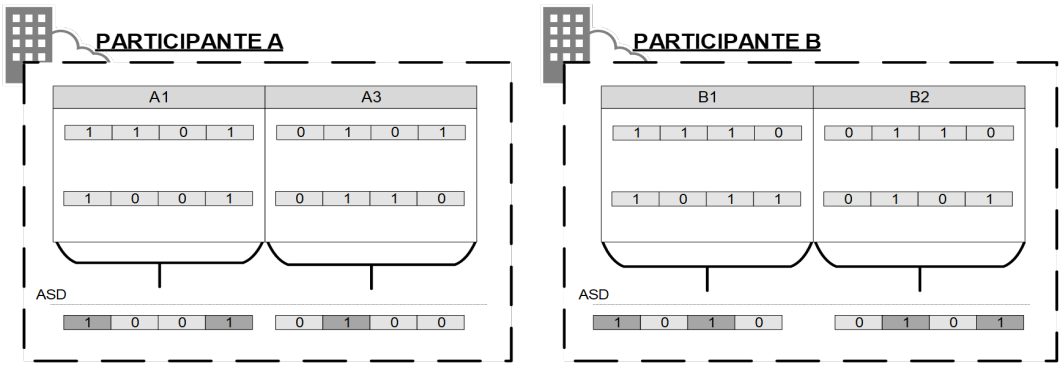


Figura 4.6: Processo de criação da ASD

Para computar a similaridade das ASD é utilizada a função *Jaccard* (Equação 4.6) sobre os valores dos *MinHash* (*array de bits*). A Figura 4.7 ilustra o cálculo da similaridade entre

os atributos A1, A3, B1 e B2, onde é possível constatar que os atributos nome (A1 e B1) e data de nascimento (A3 e B2) apresentaram os maiores valores de similaridade.

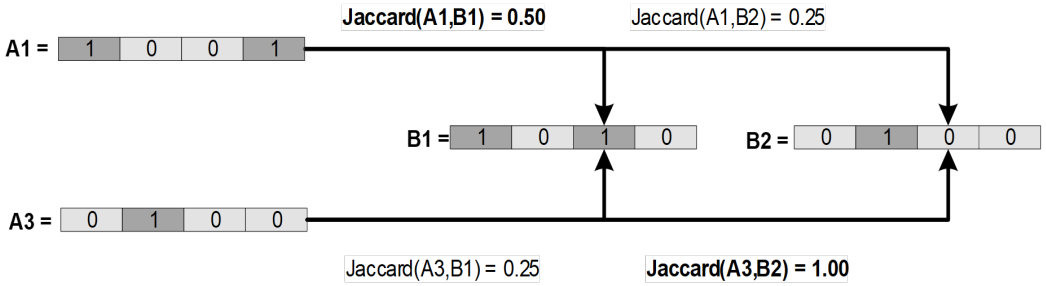


Figura 4.7: Comparação das ASD.

Na seção seguinte será apresentado como o PAC identifica os atributos similares com base das assinaturas geradas pelos participantes.

4.3.4 Pareamento Privado de Atributos

Para melhor ilustrar esta fase, o exemplo da Figura 4.3 foi redesenhado na Figura 4.8 com a supressão da fase de Padronização e Amostragem, e com o detalhamento das fases de Criação das Assinaturas e do Pareamento Privado de Atributos.

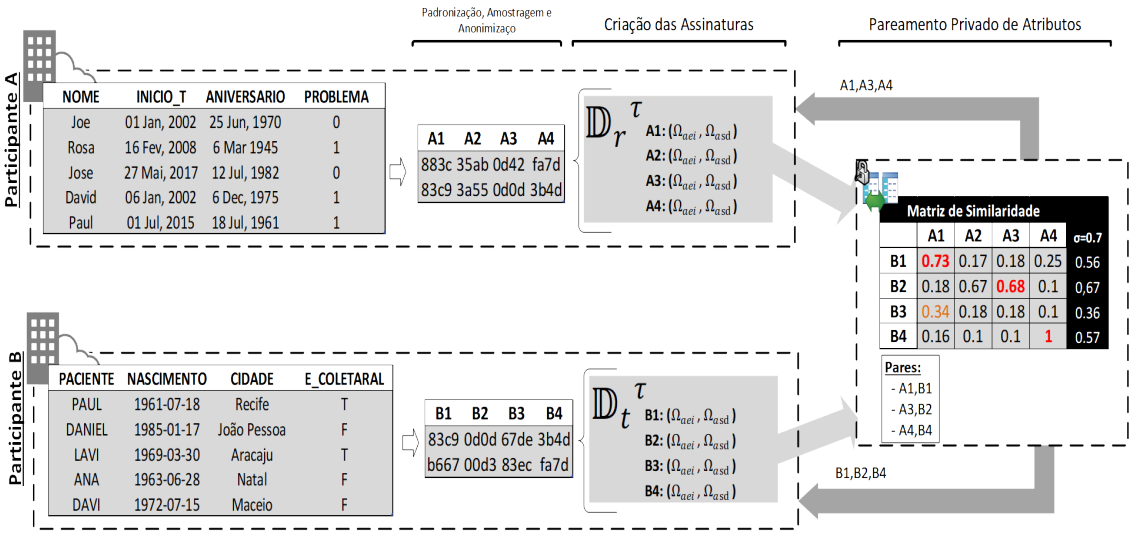


Figura 4.8: Detalhamento da fase de Pareamento Privado de Atributos do PAC.

Como ilustrado na Figura 4.8, cada participante gera um par de Assinaturas (Ω_{aei}, Ω_{asd}) para cada atributo de sua base de dados. Estas Assinaturas são enviadas para a Terceira Parte Confiável (TPC) que, por sua vez, executa o Pareamento Privado de Atributos (PPA),

obedecendo o modelo de segurança honesto, porém curioso. Por sua vez, a TPC constrói uma matriz de similaridade, $M_{i\times j} = \begin{bmatrix} m_{1,1} & \cdots & m_{1,j} \\ \vdots & \vdots & \vdots \\ m_{i,1} & \cdots & m_{i,j} \end{bmatrix}$, utilizando as Assinaturas dos atributos enviadas pelos participantes. Assim, assuma que $\Omega(\mathfrak{a}^\tau)$ represente a assinatura do atributo $\mathfrak{a}^\tau \in \mathbb{D}^\tau$. Desse modo, para calcular os elementos da matriz de similaridade ($m_{i,j}$) das duas bases de dados $\mathbb{D}_r^\tau$ e $\mathbb{D}_r^\tau$ o PPA utiliza a Equação 4.7, onde $(i,j)|i \in \mathbb{D}_r^\tau ej \in \mathbb{D}_t^\tau$.

$$m_{i,j} = w_{aei} \times (H\_sim(\Omega_{aei}[i], \Omega_{aei}[j])) + w_{asd} \times (J\_mh(\Omega_{asd}[i], \Omega_{asd}[j])) \tag{4.7}$$

A utilização de pesos (w_{aei} e w_{asd}) na Equação 4.7 permite que uma assinatura seja priorizada sobre a outra, esta priorização permite que a TPC realize um ajuste fino em alguns casos de maior importância. Como cada elemento $m_{i,j}$ da Matriz representa o valor de similaridade entre os atributos i e j, o somatório dos pesos atributos para assinaturas não deve ultrapassar 1 ($w_{aei} + w_{asd} = 1$), assegurando que o valor de similaridade final dos atributos esteja contido em um intervalo entre "0" e "1", representado dissimilaridade e similaridade máxima, respectivamente. O Capítulo 5 apresenta uma investigação sobre os valores dos pesos a serem utilizados para diferentes conjuntos de dados.

Para apresentar o algoritmo que o PPA utiliza para identificar o pareamento dos atributos é necessário introduzir dois conceitos: i) o v*alor mínimo de similaridade* (ζ); e ii) o *máximo valor global*. O valor mínimo de similaridade (ζ) é um valor calculado para cada linha da matriz ($M_{i\times j}$) utilizando um limiar escolhido pela TPC (σ) para indicar quanto ζ deve ser maior que a média dos valores de similaridade da linha (i) de $M_{i\times j}$. Ao utilizar ζ, o algoritmo assegura que o valor de similaridade mínimo entre dois atributos deve ser maior que a média em $\sigma\%$. Para melhor ilustrar este conceito, a Matriz de similaridade construída com os dados do exemplo da Figura 4.8 é exibida em detalhe na Figura 4.9.

Matriz de Similaridade					σ=0.7	
	A1	A2	A3	A4	média	ζ
B1	0.73	0.17	0.18	0.25	0.33	0.56
B2	0.18	0.67	0.68	0.1	0.39	0,67
B3	0.4	0.18	0.18	0.1	0,21	0.36
B4	0.16	0.1	0.1	1	0,34	0.57

Figura 4.9: Matriz de similaridade exibindo uma coluna com o valor médio de similaridade das linhas (média) e com o valor mínimo de similaridade (ζ)

A média dos valores de similaridade da primeira linha da matriz (destacada em vermelho) é igual a 0.33. Ao considerar um limiar de 70% ($\sigma = 0.7$), ou seja, 70% maior que média da linha (1.7×0.33), para definir ζ, a TPC selecionará os atributos que são 70% maiores que a média da linha, ou seja, no exemplo $\zeta = 1.7 * 0.33 \rightarrow 0.565$.

Por sua vez, o *máximo valor global* certifica que o elemento $m_{i,j}$ tenha o maior valor da linha e da coluna da matriz, assegurando que cada atributo esteja em apenas um par (pareamento 1:1). A utilização do *máximo valor global* é ilustrado na terceira linha (destacada em azul) da Figura 4.9. Observe que o elemento $m_{1,3}$ (A1, B3) tem um valor superior ao valor mínimo de similaridade (0.36), contudo ele não foi pareado pois o maior valor da coluna é 0.73 ($m_{1,1}$).

A tarefa de paramento privado de atributos é descrita no Algoritmo 1, que recebe como parâmetros de entrada $M_{i \times j}$ e σ. Inicialmente, o algoritmo itera sobre as linhas da matriz $M_{i \times j}$, onde, para cada linha, é calculado ζ (linhas 3 a 6 do Algoritmo 1). Em seguida, é selecionado o atributo (representado pelo índice da coluna) atendendo a três requisitos: i) tenha o maior valor de similaridade da linha; ii) possua o valor de similaridade maior que ζ; e iii) represente o máximo valor global do atributo (linhas 7 a 15 do Algoritmo 1). Por fim, o algoritmo retorna uma lista com os pares de atributos pareados (linha 16 do Algoritmo 1).

Algoritmo 1: Algoritmo de Pareamento Privado de Atributos (PPA)

Entrada: $M_{i\times j}$,σ

Saída : $\Gamma_{\mathbb{D}_r^\tau,\mathbb{D}_t^\tau}$

$\Gamma_{\mathbb{D}_r^\tau,\mathbb{D}_s^\tau} \leftarrow []$
for $linha$ *in* $M_{i\times j}$ **do**
  $media\_linha \leftarrow media(linha)$
  $max\_linha \leftarrow max(linha)$
  $coluna \leftarrow getColuna(M, linha, max\_linha)$
  $\zeta \leftarrow media\_linha + (media\_linha \times \sigma)$
  if $line\_max\_val \geq \zeta$ **then**
    $atributo\_1 \leftarrow getAttibuto(linha, max\_linha)$
    $col\_max\_val \leftarrow max(coluna)$
    $atributo\_2 \leftarrow getAtributo(coluna, col\_max\_val)$
    if $max\_linha > col\_max\_val$ **then**
      $\Gamma_{\mathbb{D}_r^\tau,\mathbb{D}_s^\tau} \leftarrow \Gamma_{\mathbb{D}_r^\tau,\mathbb{D}_s^\tau} + \langle atributo\_1, atributo\_2\rangle$
    end
  end
end
return $\Gamma_{\mathbb{D}_r^\tau,\mathbb{D}_s^\tau}$

Após a execução do PPA, a terceira parte confiável envia para cada participante uma lista que indica quais dos seus atributos devem ser utilizados nas etapas seguintes da REPP. Desse modo, os participantes podem executar uma tarefa de REPP sem a necessidade de divulgar informações sobre os dados ou o esquema.

4.3.5 Detalhamento do Protocolo utilizado pelo PAC

Como apresentado nas seções anteriores, o PAC é uma abordagem que utiliza um protocolo com três participantes (*three-party protocol*) e considera um modelo de adversário honesto porém curioso. Assim, para um melhor entendimento, esta seção apresenta o detalhamento das mensagens trocadas pelos participantes durante a execução do PAC.

A Figura 4.10 ilustra o fluxo de mensagens entre os participantes e a Terceira Parte

Confiável (TPC) durante a execução do PAC. No primeiro momento, antes de iniciar a etapa de Apresentação, os participantes concordam em relação à entidade a ser utilizada na tarefa de REPP e o formato padrão dos dados a ser usado durante a fase de Padronização do PAC. Em seguida, é iniciada a etapa de Padronização (com pré-processamento) que computa estatísticas para a etapa de Anonimização (e.g. número de bits a serem utilizados no Filtro de Bloom), converte os dados para o formato padrão e os armazena em um esquema de tabela única. Em seguida, os participantes trocam informações sobre os parâmetros a serem utilizados na etapa de Anonimização de dados. Note que, até esse momento, as informações foram compartilhadas apenas entre os participantes (A e B), ou seja, a TPC não conhece o formato padrão ou os parâmetros de anonimização, dificultando a reidentificação do dado original pela TPC.

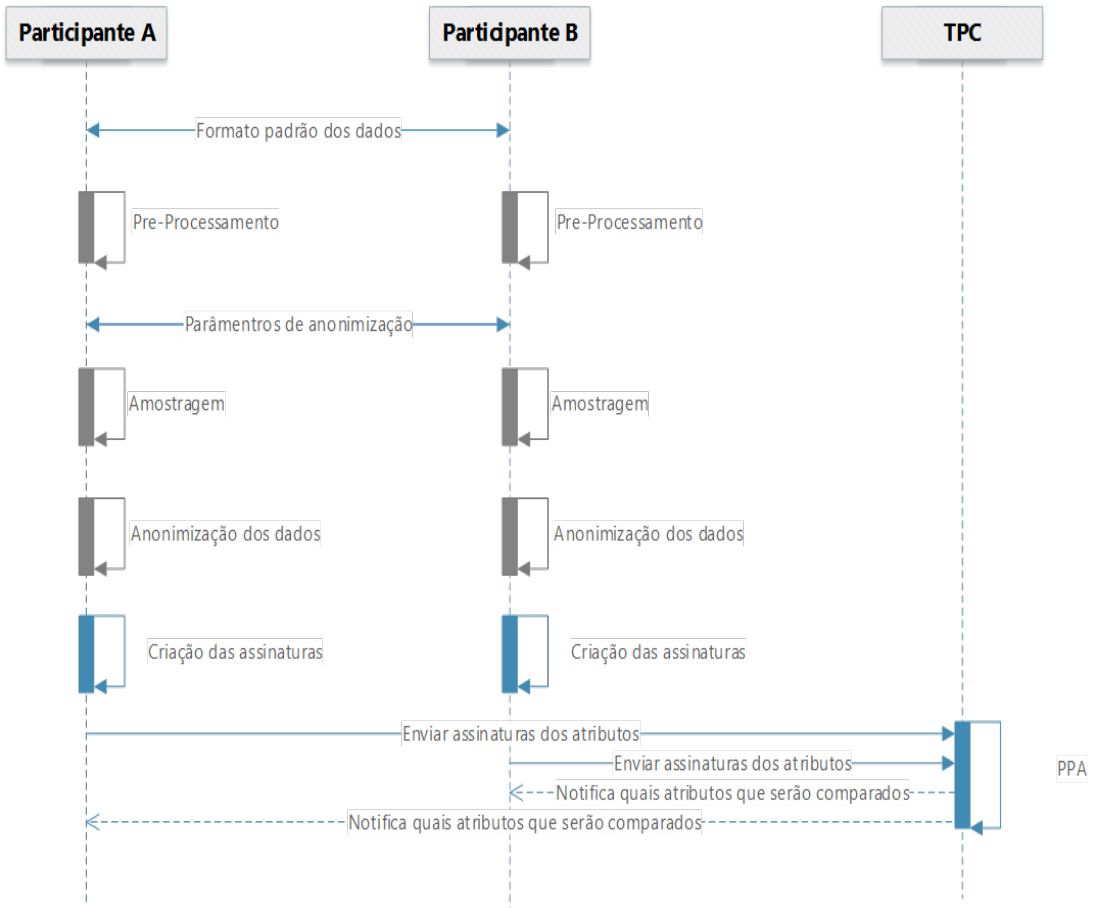


Figura 4.10: Detalhamento da troca de mensagens durante a execução do PAC.

Por fim, as assinaturas são geradas e enviadas para a TPC que executa o PPA e, ao final, notifica os participantes sobre quais dos seus atributos devem ser utilizados na REPP.

Observe que apenas nessa etapa a TPC recebe mensagens (com as assinaturas) dos participantes, e que, após a execução do PPA, a TPC notifica, a cada participante, quais dos seus atributos devem ser utilizados. Ao enviar individualmente os atributos que devem compor o QID, a utilização do PAC impede que um participante aprenda sobre os dados (atributos) dos demais participantes.

4.4 Considerações Finais

Nesta seção foi formalizada e detalhada uma abordagem para a etapa de Apresentação da REPP. No capítulo seguinte será apresentada a validação (da eficácia, eficiência e privacidade) do Pareamento às Cegas de Atributos.

Capítulo 5

Validação e Experimentos

Neste capítulo são apresentados os experimentos conduzidos para validar empiricamente a abordagem de Pareamento de Atributos às Cegas (PAC). Os experimentos apresentados têm o objetivo de avaliar a eficiência e eficácia do PAC no que se refere à identificação de atributos que as entidades de duas bases de dados possuem em comum, preservando a privacidade dos dados originais.

5.1 Bases de Dados Utilizadas

Os experimentos são conduzidos em oito bases de dados reais, divididas em quatro cenários. Cada cenário consiste de um par de bases de dados e explora características que dificultam o processo de identificação de atributos, tais como dados sujos e número de atributos e entidades desproporcionais. Os cenários são descritos abaixo:

i Eleitores: este cenário utiliza bases de dados com informações sobre eleitores (nome, endereço, filiação partidária, data de nascimento, entre outras) de dois estados norte-americanos, Carolina do Norte (NCVR) [53] e Ohio (OHVR) [54];

ii Restaurantes: este cenário utiliza duas bases de dados com informações sobre restaurantes (nome, endereço, especialidade, faixa de preço, entre outros) dos Estados Unidos. Neste cenário foram utilizadas bases de dados do TripAdvisor[1] e Yelp [74], duas plataformas onde os dados são inseridos pelos próprios usuários (*crowdsourcing*),

[1]Dados coletados por um crawler disponível no endereço http://github.com/thiagonobrga/

ou seja, os dados destas bases estão propensos a erros de digitação e informações faltantes, uma vez que são os usuários que introduzem dados nestas bases;

iii Medicamentos: este cenário contém bases de dados médicos com informações sobre reações adversas a medicações (droga, dose, reação, data, forma, entre outros). São utilizadas bases provenientes da *Federal Drug Administration* (FDA) dos Estados Unidos [67] e da *Marketed Health Products Directorate* (MHPD) do Canadá [12];

iv Servidores Públicos: neste cenário é utilizada uma base de dados de servidores públicos (nome, cpf, matricula, data de nomeação, órgão de lotação, entre outros) que recebem salário da União e uma base de servidores federais com punições administrativas (número do processo, nome, cpf, data da punição, motivo, entre outros). Para tal, são utilizados os dados dos servidores federais disponibilizados pelo Ministério do Planejamento, Orçamento e Gestão (MPOG) e das punições dos servidores públicos federais disponibilizados pela Controladoria Geral da União (CGU) [16]. Para atender a requisitos legais o MPOG e a CGU ofuscam alguns atributos, por exemplo, três dígitos do CPF são substituídos por * (e.g. 759.***.231-14). Tal alteração (ofuscamento) em alguns atributos torna mais complexo o processo de pareamento privado de atributos.

Os cenários descritos anteriormente representam diferentes entidades (eleitores, restaurantes, servidores públicos e reações adversas de medicamentos) comumente utilizadas em diferentes áreas de conhecimento (Medicina, Direito, Administração Pública, Urbanismo, entre outras). Para cada cenário foi solicitado a um especialista que indicasse as correspondências (gabarito) entre atributos das duas bases de cada cenário. O detalhamento das bases (número de atributos, número de entidades e quantidade de atributos correspondentes nas bases de dados) estão listados na Tabela 5.1.

Tabela 5.1: Detalhamento das bases de dados utilizadas.

#	Cenário	Base de Dados	País	Número de Atributos	Número de Entidades	Número de Correspondências entre atributos
1	Eleitores	ncvr	US	16	7.696.365	7
		ohvr	US	10	7.898.805	
2	Restaurantes	trip	US	12	115.369	7
		yelp	US	16	37.623	
3	Medicamentos	fda	US	10	12.324	7
		cvp	CA	19	499.134	
4	Servidores Públicos	mpog	BR	26	321.480	5
		cgu	BR	12	4.947	

A utilização de bases de dados reais nos cenários busca retratar potenciais problemas que o PAC pode enfrentar. Dentre os problemas destacam-se: dados sujos e faltantes, desbalanceamento no número de atributos e correspondências, além de bases de dados grandes contendo milhões de entidades. É importante ressaltar que não foram utilizadas bases de dados com um maior número de atributos pois, de acordo com o nosso melhor conhecimento, os trabalhos que relatam casos de uso da REPP [39, 43, 57, 62, 68, 75] utilizam entidades com no máximo 20 atributos [68].

5.2 Métricas de Qualidade

A fim de avaliar a qualidade (eficácia) do pareamento dos atributos foram utilizadas três métricas (Precision, Recall e F-measure), tipicamente utilizadas na área de Recuperação de Informação [13]. Para adaptar estas medidas ao nosso contexto é preciso formalizar alguns conceitos. Dadas duas bases de dados, assuma que o **E** representa o conjunto dos pares de atributos similares (correspondentes) existentes nas duas bases, e que **R** representa o

conjunto de pares de atributos classificados como similares. Desse modo, as métricas são definidas como:

$$Precision = \frac{|E \cap R|}{|R|} \tag{5.1}$$

$$Recall = \frac{|E \cap R|}{|E|} \tag{5.2}$$

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{5.3}$$

Intuitivamente, um valor de Precision baixo indica que correspondências falsas foram identificadas pela abordagem, enquanto um Recall baixo implica que a abordagem não classificou alguns pares de atributos, ou seja, deixou de identificar atributos. Já a F-measure (F1) combina a Precision e o Recall em uma média harmônica, sendo bastante utilizada para indicar o balanceamento entre as duas métricas anteriores.

5.3 Experimentos e Hipóteses

A validação da abordagem de pareamento de atributos é realizada em quatro avaliações, na qual a primeira avaliação tem como objetivo avaliar a eficácia (qualidade) do PAC, considerando a utilização de todas as entidades (presentes nas bases de dados) e as estratégias de amostragem (Independente e Conjunta). A segunda avaliação, por sua vez, investiga o impacto da atribuição de pesos diferentes para as assinaturas na composição da matriz de similaridades (apresentada na Seção 4.3.4), seguida pela avaliação da privacidade. Por fim é realizada a avaliação da eficiência do pareamento dos atributos. A seguir, são listadas as hipóteses para cada avaliação:

Avaliação da Qualidade

H_1 : No contexto de REPP é possível identificar as correspondências entre atributos sem a necessidade de os participantes divulgarem informações que possam ser utilizadas para quebrar a privacidade dos dados originais?

H_2 : No contexto do PAC, a utilização de amostras de entidades produz resultados significativamente diferentes quando comparados com a utilização de todas as entidades das bases de dados?

H_3 : No contexto do PAC, a estratégia utilizada para escolher o número de elementos da amostra (número de entidades) tem impacto significativo nos resultados?

H_4 : No contexto do PAC, amostras com diferentes números de elementos produzem resultados significativamente diferentes quanto comparadas entre si?

Avaliação das Assinaturas

H_5 : No contexto do PAC, é necessária a utilização de duas assinaturas de dados (ASD e AEI)?

H_6 : No contexto do PAC, a atribuição de pesos diferentes às assinaturas de dados (AEI e ASD) para calcular os elementos da matriz de similaridade (Equação 4.7) produz resultados de qualidade significativamente diferentes?

Avaliação de Privacidade

H_7 : O PAC é capaz de preservar a privacidade de dados e esquemas utilizados em uma tarefa de REPP?

Avaliação da Eficiência

H_8 : No contexto da REPP, qual o custo computacional da utilização do PAC?

Nas seções seguintes são apresentados e discutidos os resultados de cada avaliação supracitada.

5.4 Avaliação da Qualidade

Nesta seção é apresentada a investigação empírica da capacidade do PAC em parear os atributos (H_1), das metodologias do PAC para calcular o tamanho das amostras (H_2 e H_3) e do impacto do número de elementos das amostras para verificar as hipóteses (H_4).

5.4.1 Desenho Experimental

O desenho experimental desta avaliação constitui em: i) extrair amostras de cada base de dados variando o número de entidades de cada amostra e a estratégia para escolha do número de elementos da amostra (Conjunta ou Independente), e ii) utilizar essas amostras como entrada para o PAC, variando o limiar de classificação.

O desenho experimental é sumarizado na Tabela 5.2, este é composto por cinco colunas onde as duas primeiras apresentam o cenário e as bases de dados utilizadas nos experimentos. A terceira coluna do quadro apresenta a estratégia utilizada para definir o número de elementos da amostra (descritas no Capítulo 4). O valor "*Independente*" representa a estratégia na qual cada participante extrai uma amostra com o mesmo percentual (números de elementos distintos), o valor "*Conjunta*" representa a estratégia na qual cada participante envia para a terceira parte confiável (TPC) o número de entidades presentes em sua base de dados para que a TPC determine o número de elementos da amostra considerando o tamanho das duas bases de dados. A quarta coluna representa o número de elementos das amostras, nestes experimentos foram utilizados os valores 0,25%, 0,5%, 1%, 1,5% e 2% para ambas as estratégias. Por fim, a coluna Limiar de Classificação representa o limiar (α) utilizado pelo PPA (Algoritmo 1).

Tabela 5.2: Desenho Experimental da Avaliação da Qualidade.

Cenário	Base de Dados	Estratégia	Número de Elementos da Amostra					Limiar de Classificação
			0,25%	0,50%	1%	1,50%	2%	
Eleitores	-	Conjunta	38.988	77.976	155.952	233.928	311.903	[0,1, ... ,0,9]
	ncvr	Independente	19.241	38.482	76.964	115.455	153.927	[0.1, ... ,0.9]
	ohvr		19.747	39.494	78.988	118.482	157.976	[0.1, ... ,0.9]
Restaurantes	-	Conjunta	382	765	1.530	2.295	3.060	[0.1, ... ,0.9]
	trip	Independente	288	577	1.154	1.731	2.307	[0.1, ... ,0.9]
	Yelp		94	188	376	564	752	[0.1, ... ,0.9]
Medicamentos	-	Conjunta	1.279	2.557	5.115	7.672	10.229	[0.1, ... ,0.9]
	fda	Independente	31	62	123	185	246	[0.1, ... ,0.9]
	mhdp		1.248	2.496	4.991	7.487	9.983	[0.1, ... ,0.9]
Servidores Públicos	-	Conjunta	816	1.632	3.264	4.896	4.947	[0.1, ... ,0.9]
	mpog	Independente	804	1.607	3.215	4.822	6.430	[0.1, ... ,0.9]
	cgu		12	25	49	74	99	[0.1, ... ,0.9]

Com o intuito de reduzir o viés introduzido pelo processo de amostragem (randômico) foram extraídas cinco amostras para cada combinação das colunas "Estratégia para Escolha do Número de Elementos da Amostra" e "Base de dados". Em outras palavras, para cada linha da Tabela 5.2, foram extraídas cinco amostras (distintas) para cada valor presente na coluna "Número de elementos da Amostra". Para investigar se o PAC[2] é capaz de parear atributos em um contexto de REPP (H_1), os resultados alcançados pela abordagem proposta foram comparados com o resultado de um competidor (DUMAS [8]) e com a utilização de todas as entidades das bases de dados.

Os experimentos desta seção foram executados em uma instância Linux 4.0.1 do Microsoft Azure (DS14-4_v2 Standard) com 4 CPUs (Intel Xeon® E5-2673 v3 de 2,4 GHz) e 112 GB de memória RAM. É importante ressaltar que nestes experimentos foram atribuídos pesos iguais para as assinaturas ($w_{aei} = 0.5$ e $w_{asd} = 0.5$). Os parâmetros de anonimização utilizados em cada cenário são expostos no Apêndice A.

5.4.2 Competidor

O competidor escolhido para esta investigação foi o DUMAS proposto por Bilke e Naumann [8], este competidor utiliza as entidades duplicadas (em ambas as bases) para realizar o pareamento dos atributos (como apresentado no Capítulo 3). A comparação do PAC com o DUMAS pode não parecer uma comparação justa, pois o DUMAS não considerar a privacidade dos dados durante o processo, no entanto, uma maneira de avaliar a eficácia do PAC é compara-lo com uma abordagem tradicional. Outra razão para utilizar o DUMAS é que, assim como o PAC, ambas as abordagens utilizam um esquema de tabela única com os atributos opacos, ou seja, as abordagens não consideram informações sobre o esquema da base de dados (nome de atributos, tipo de atributos, entre outras), apenas os valores dos atributos das entidades para identificar os atributos.

Os experimentos foram executados utilizando uma versão DUMAS[3] disponibilizada pelos autores. Os parâmetros e as bases dados foram preparados e utilizados de acordo com a documentação disponibilizada pelos autores.

[2]Disponível em https://github.com/thiagonobrega/BAP

[3] https://hpi.de/naumann/projects/repeatability/algorithms/dumas-duplicate-based-matching-of-schemas.html

5.4.3 Resultados

Os resultados dos experimentos são ilustrados graficamente. Para cada métrica de qualidade é plotado um gráfico, no qual o eixo vertical representa a métrica de qualidade e o eixo horizontal o limiar de classificação utilizado para definir se os atributos são similares. Em razão da utilização de amostras com diferentes números de elementos cada estratégia tem o seu valor médio ilustrado por uma linha colorida envolta por uma silhueta (*ribbon*) de mesma coloração, a qual representa o erro padrão para cada limiar. Por sua vez, o DUMAS é representado por uma linha horizontal tracejada na cor vermelha.

As Figuras 5.1, 5.2, 5.3 e 5.4, as quais são compostas por três gráficos, sumarizam os resultados alcançados pelas diferentes combinações dos parâmetros expostos na Tabela 5.1 para os cenários Eleitores, Restaurantes, Medicamentos e Servidores Públicos, respectivamente. A Figura 5.1 ilustra os resultados de eficácia (qualidade) para o cenário Eleitores, no qual a abordagem DUMAS consumiu por completo os 112GB de RAM disponíveis e não conseguiu finalizar a execução por falta de memória RAM. Em razão disto foi considerado que esta abordagem não identificou qualquer atributo.

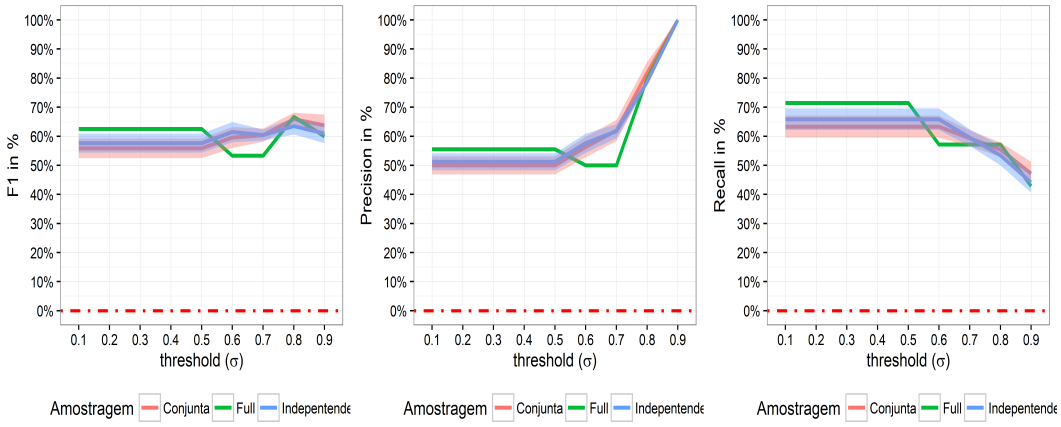


Figura 5.1: Resultados de Qualidade para o cenário **Eleitores**.

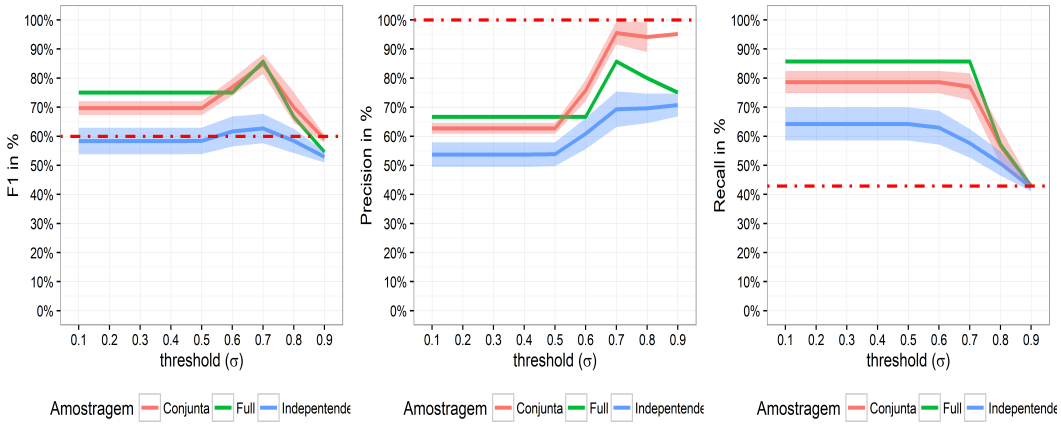


Figura 5.2: Resultados de Qualidade para o cenário **Restaurantes**.

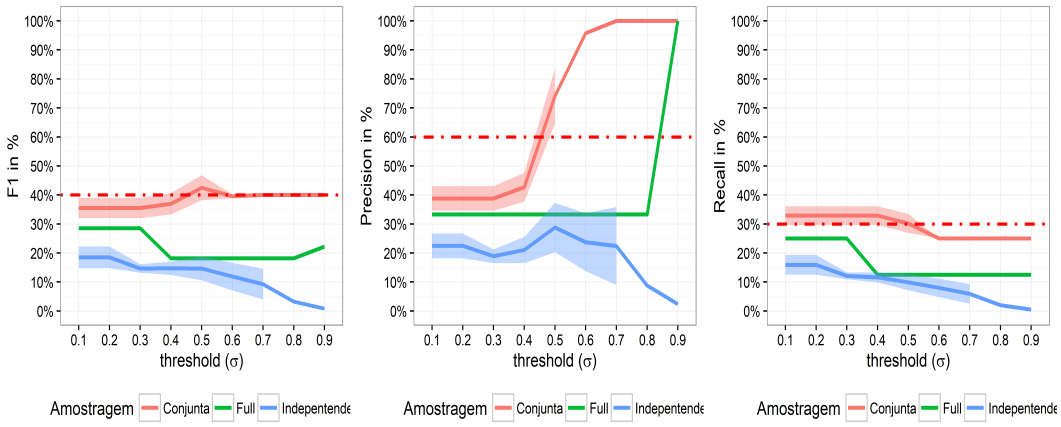


Figura 5.3: Resultados de Qualidade para o cenário **Medicamentos**.

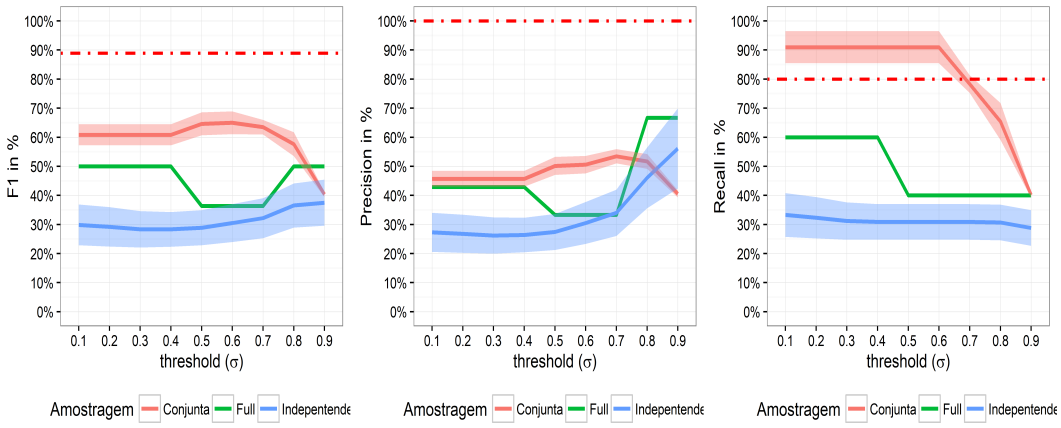


Figura 5.4: Resultados de Qualidade para o cenário **Servidores Públicos**.

5.4.4 Discussão

Observando as Figuras 5.1 a 5.4 é possível constatar que a utilização da estratégia Conjunta, para selecionar o número de elementos das amostras, obteve resultados de eficácia superiores para quase todos os cenários, com exceção do cenário Eleitores, no qual a estratégia Conjunta demostrou resultados similares aos da estratégia Independente e com a utilização de todos os registros da base de dados (Full).

Por sua vez, ao comparar os resultados alcançados pela estratégia Independente com a estratégia Conjunta e a utilização de todas as entidades percebe-se que esta estratégia Independente alcançou resultados inferiores para todos os casos (com exceção do cenário Eleitores, no qual obteve um resultado similar ao da estratégia Conjunta). Este fato pode ser explicado pelo fato da estratégia Independente utilizar amostras com números de elementos distintos nas amostras de cada participante (ilustrado no Tabela 5.2). A utilização de amostras com números de elementos distintos faz com que a assinatura AEI gere valores de Entropia diferentes para atributos similares.

Para melhor explicar os resultados, os cenários serão apresentados de acordo com o desbalanceamento entre o número de entidades e o número de atributos nas suas bases de dados originais. No cenário Eleitores (Figura 5.1), no qual existe uma equivalência entre o número de entidades e atributos, é possível constatar estatisticamente (utilizando o teste de Wilcox com confiança de 95%) que as estratégias Independente, Conjunta e Full apresentam resultados similares. Este fato é explicado pelo equilíbrio entre o número de entidades das bases de dados (7.6 milhões de entidades na NCVR e 7.8 milhões na OHVR). Desse modo, as amostras extraídas e as bases de dados apresentam um número equivalente de entidades (descrito na Tabela 5.2).

No cenário Restaurantes (Figura 5.2), no qual o desbalanceamento entre o número de entidades é leve (com uma razão $\frac{1}{4}$ entre o número de entidades entre a menor e a maior base de dados), é possível constatar que os resultados de eficácia obtidos ao utilizar estratégia Conjunta são estatisticamente equivalentes (com 95% de confiança) aos resultados alcançados ao considerar todas as entidades da base. Ao comparar os resultados da estratégia Conjunta com o DUMAS é possível constatar que os resultados do PAC superam o DUMAS em termos de F-measure e Recall para todos os limiares testados. Por sua vez, a métrica Precision tem o mesmo resultado do DUMAS (100%) quando utilizados os limiares de 0.7 e 0.8.

Por sua vez o cenário Medicamentos (Figura 5.3), no qual o desbalanceamento entre o número de entidades é elevado (com uma razão $\frac{1}{40}$ entre o número de entidades em relação a menor e maior bases de dados), a estratégia Conjunta tem resultados de eficácia superiores em todas as métricas quando comparada à utilização de todas as entidades da base de dados. Ao comparar o resultado da estratégia Conjunta com o DUMAS é possível perceber que: i) os resultados obtidos com as métricas F-measure e Recall do PAC e a do DUMAS são estaticamente equivalentes (com 95% de confiança), ii) e o resultado de Precision da PAC foi superior ao DUMAS a partir do limiar 0.5. Este resultado já era esperado, pois a utilização da estratégia Conjunta pelo PAC faz com que as assinaturas alcance melhores resultados que as demais por: i) as assinaturas serem criadas a partir de amostras com o mesmo número de elementos, e ii) por reduzir o número de entidades com erros de digitação (para a ASD), uma vez que esta base apresenta informações inseridas por profissionais de saúde do Canada e dos Estados Unidos.

Por fim, o cenário Servidores Públicos (Figura 5.4), o qual representa o maior desafio para o PAC por apresentar as seguintes características: i) elevado desbalanceamento entre o número de entidades (com uma razão $\frac{1}{60}$ entre o número de entidades presente na menor e maior base de dados), ii) maior número de atributos (26 atributos na base de dados MPOG e 16 atributos na base CGU), e iii) menor número de correspondências de atributos (apenas cinco atributos estão presentes nas duas bases de dados). Neste cenário, a utilização da estratégia Conjunta propiciou, considerando o limiar de 0.6, uma importante melhora de mais de 50% no Recall e 25% na Precision. Contudo, o PAC não foi capaz de superar os resultados alcançados pelo DUMAS. Este fato pode ser explicado pelo fato do DUMAS necessitar de poucas entidades duplicadas para realizar o pareamento, enquanto o PAC teve seu desempenho prejudicado em razão do: i) mascaramento dos dados, e ii) desbalanceamento no número de atributos e entidades das bases de dados.

Os resultados dos experimentos (Figuras 5.1 a 5.4) dão suporte às Hipóteses H_1 (o PAC foi capaz de parear atributos em um contexto de REPP), H_2 (a utilização de amostras aumenta a eficácia do PAC) e H_3 (a utilização da estratégia Conjunta para definir o número de elementos das amostras apresentou um impacto significativo nos resultados).

Para investigar o impacto da utilização de amostras de diferentes tamanhos, com 0.25%, 0.5%, 1%, 1.5% e 2% do número total de entidades presentes nas bases de dados, foram

considerados apenas os resultados do PAC obtidos aplicando a estratégia Conjunta, uma vez que esta apresentou uma eficácia superior aos da estratégia Independente. Os resultados (F-measure, Precision e Recall) são apresentados na Figura 5.5. Para cada métrica é plotado um gráfico no qual o eixo vertical representa a métrica de qualidade e o eixo horizontal o limiar de classificação utilizado no Pareamento Privado de Atributos. Os resultados alcançados pelo PAC têm o seu valor médio representado por uma linha colorida envolta por uma silhueta (*ribbon*) de mesma coloração, a qual representa o erro padrão para cada limiar. Por sua vez, o resultado da utilização de todas as entidades é representado por uma linha na cor preta.

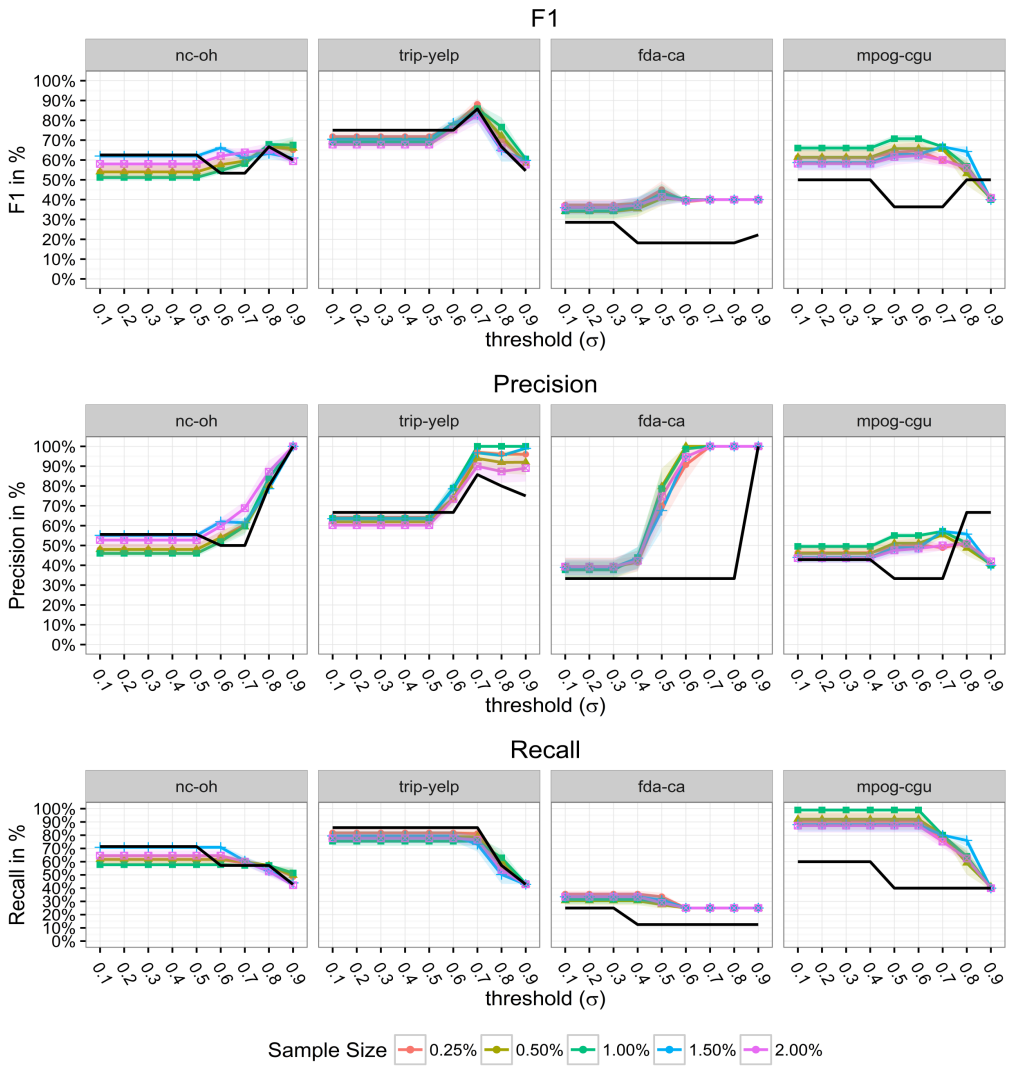


Figura 5.5: Resultados alcançados pelo PAC (com a estratégia Conjunta) para amostras de diferentes tamanhos.

Os resultados expostos na Figura 5.5 demonstram que, para todos os fatores avaliados

(cenários, métricas, limiares e números de elementos das amostras), o fator que resultou em maiores ganhos de qualidade foi a escolha de valores de limiares entre 0.6 e 0.8. Assim, em razão da dificuldade de identificar (visualmente) qual amostra obteve os melhores resultados, foi utilizado um teste estático (Análise de Componente Principal) para selecionar a amostra que teve o maior impacto na eficácia (avaliada pela F-measure). Este teste demonstrou que amostras com 1.5% e 2% do número de entidades totais das bases de dados são mais relevantes que os demais valores (0.25%, 0.5% e 1%) quando considerado a F-measure. Desse modo, foi realizado um teste de hipótese com as seguintes hipóteses:

H_0 Neste experimento não há diferença da F-measure ao se alterar o número de elementos das amostras;

H_a Neste experimento há uma diferença ao se utilizar amostras com 1.5% e 2%.

O resultado do teste de hipótese retornou um *p-valor* de 0.00476. Desse modo, a hipótese nula foi rejeitada e confirmada a hipótese que considera que o número de entidades influencia no resultado final do PAC (H_4).

Em resumo, os experimentos apresentados nesta seção confirmaram as quatro hipóteses iniciais. Isso mostra que o PAC é capaz de parear atributos com uma qualidade equivalente a uma abordagem que não considera a privacidade dos dados. Além disso, a estratégia e o número de elementos tem impacto sobre a qualidade do pareamento.

5.5 Avaliação das Assinaturas

Nesta seção é investigado o peso atribuído às assinaturas AEI e ASD para a construção da Matriz de Similaridade utilizada pelo Algoritmo de Pareamento Privado de Atributos. Desse modo, pretende-se avaliar se a utilização de uma média ponderada dos valores de similaridade das Assinaturas (ASD e AEI) para a construção da Matriz de Similaridades possui resultados significativos na eficácia do PAC, ou seja, se a atribuição de pesos diferentes paras as assinaturas pode melhorar a eficácia do PAC.

5.5.1 Desenho Experimental

Como exposto na Seção 4.3.4, o algoritmo de Pareamento Privado de Atributos (Algoritmo 1) recebe como entrada uma Matriz de Similaridades na qual seus elementos ($m_{i,j}$) são calculados pela Equação 4.7, transcrita a seguir, onde w_{aei} é o peso da AEI e w_{asd} o peso do ASD.

$$m_{i,j} = w_{aei} \times (H\_sim(\Omega_{aei}[i], \Omega_{aei}[j])) + w_{asd} \times (J\_mh(\Omega_{asd}[i], \Omega_{asd}[j]))$$

O desenho deste experimento utilizou as amostras extraídas usando a estratégia Conjunta (estratégia que apresentou o melhor resultado no experimento anterior) para avaliar o impacto (considerando as métricas de qualidade) da atribuição de diferentes pesos para as assinaturas ASD e AEI para gerar a Matriz de Similaridades. A Tabela 5.3 ilustra a variação dos parâmetros utilizados nesse experimento.

Tabela 5.3: Parâmetros utilizados no Experimento.

	Peso AEI	Peso ASD	Limiar de Classificação
	0,0	1,0	0.7
Cenário	···	···	0.7
	1,0	0,0	0.7

Conforme ilustrado no Tabela 5.3, para todos os experimentos, foi utilizado o limiar que apresentou o melhor resultado nos experimentos da seção anterior (no caso, 0.7). Por sua vez, o peso (w) atribuído às assinaturas varia de 0 a 1. Em outras palavras, quando o peso atribuído a uma assinatura é 0, esta assinatura é desconsiderada e, à medida que o peso se aproxima de 1, a relevância da assinatura é aumentada. É importante observar que a soma dos pesos atribuídos não ultrapassa 1 ($w_{asd} = 1 - w_{aei}$), pois os elementos da Matriz de Similaridades devem ser normalizados entre 0 e 1. Desse modo, pretende-se investigar qual é o peso ideal para cada cenário.

5.5.2 Resultados

Os resultados deste experimento são ilustrados nos gráficos da Figura 5.6. O eixo vertical representa os valores de F-measure alcançados pelo PAC, e o eixo horizontal representa a combinação dos pesos atribuídos às assinaturas para o experimento. Os valores de F-measure alcançados são plotados em barras (com o erro padrão plotado no topo de cada barra), as quais foram coloridas de acordo com os seguintes critérios:

i controle (na cor roxa): representa o valor padrão utilizado pelo algoritmo de Pareamento Privado de Atributos (0.5 para ambas as assinaturas), ou seja, pesos iguais para as assinaturas;

ii melhor (na cor verde): indica os experimentos que alcançaram resultados estaticamente superiores aos da estratégia Controle;

iii pior (na cor azul): refere-se aos experimentos que alcançaram resultados estaticamente inferiores aos da estratégia Controle;

iv não significativo (na cor cinza): refere-se aos experimentos cujos resultados não foram estaticamente diferentes dos resultados da estratégia Controle.

Para classificar os resultados de acordo com os critérios supracitados foram executados testes de validação estática, propostos por Mayer e Butler [44], para cada combinação de pesos ilustrada na Tabela 5.3.

A Figura 5.6 ilustra os resultados obtidos com a utilização de pesos diferentes para as assinaturas. O eixo horizontal foi ordenado da esquerda para a direita reduzindo o peso da AEI e aumentado o da ASD. Desse modo, a primeira barra de cada gráfico desconsidera a ASD ($w_{aei} = 1$, $w_{asd} = 0$) e, à medida que as barras se aproximam do extremo direito do eixo horizontal, o valor atribuído à ASD é aumentado até atingir 100% ($w_{aei} = 1$, $w_{asd} = 0$).

5.5.3 Discussão

Para compreender melhor o resultado deste experimento é necessário discutir a relação entre o número de atributos totais/correspondências e o desempenho do algoritmo de Pareamento

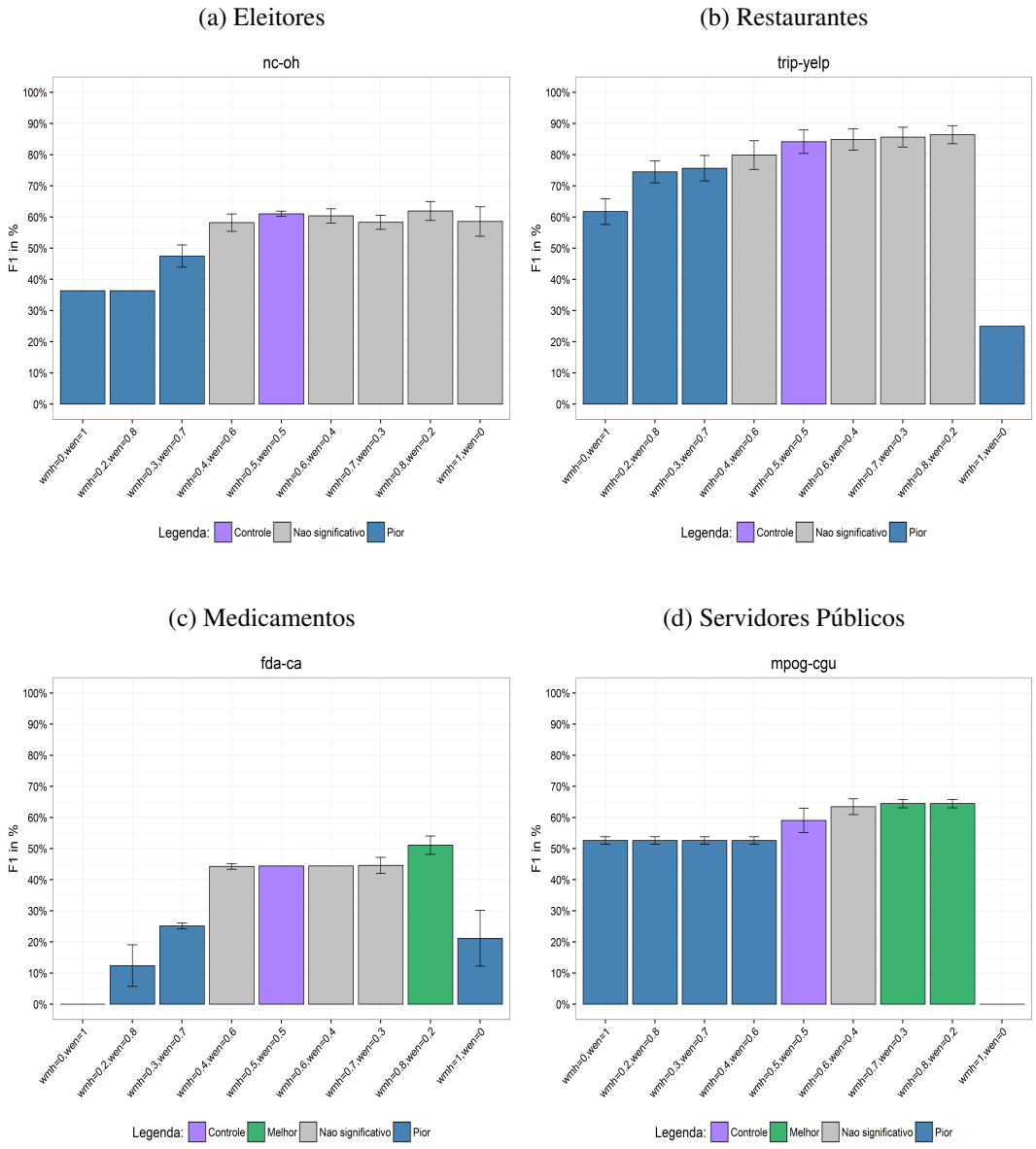


Figura 5.6: F-measure alcançada pelo PAC com a atribuição de pesos distintos para as Assinaturas.

Privado de Atributos. A Tabela 5.4 sumariza algumas informações sobre o número de atributos das bases de dados utilizadas em cada cenário. As quatro primeiras colunas foram extraídas da Tabela 5.3. Por sua vez, as colunas Número de atributos não correspondentes e Razão dos atributos representam o número de atributos não correspondentes presentes em ambas as bases de dados, e a razão entre o número de atributos das bases de dados (menor/-maior), respectivamente.

Tabela 5.4: Informações sobre as bases de dados de cada cenário.

#	Cenáro	DataSet	Número Total de Atributos	Número de Correspondências	Número de atributos não correspondentes	Razão dos atributos
1	Eleitores	ncvr	16	7	12	0,62
		ohvr	10			
2	Restaurantes	trip	12	7	14	0,75
		yelp	16			
3	Medicamentos	fda	10	7	15	0,52
		cvp	19			
4	Servidores Públicos	mpog	26	5	26	0,46
		cgu	12			

As colunas Número de atributos não correspondentes e Razão dos atributos foram apresentadas para melhor analisar o impacto que as variáveis número total de atributos e número de correspondências representam para a Matriz de Similaridades gerada pelo PPA.

A coluna Razão dos atributos tem o seu valor compreendido entre 0 e 1 e pode ser interpretada da seguinte maneira: quanto mais próximo de 1 a Matriz de Similaridades se aproxima de uma matriz quadrada; caso contrário, de uma matriz coluna. Por sua vez, o Número de atributos não correspondentes tem influência sobre o valor mínimo de similaridade (ζ), diminuindo o valor da média calculada para cada linha da Matriz de Similaridades. Em termos práticos, a utilização de matrizes com poucas colunas e muitas linhas (ou seja, quando há um número elevado de atributos não correspondentes) faz com que o PPA classifique (erroneamente) atributos que não são similares como correspondências (introdução de falsos positivos).

Em razão do exposto nos parágrafos anteriores, os cenários podem ser divididos em dois grupos: i) Eleitores e Restaurantes, os quais apresentam uma Razão de atributos próxima a 1 e números de atributos (correspondentes e não correspondes) similares; e ii) Medicamentos e Servidores Públicos, os quais apresentam uma razão de atributos menor e um maior número de atributos não correspondentes.

Ao analisar os resultados do primeiro grupo (Eleitores e Restaurantes), é possível observar que nenhum resultado superou estaticamente o resultado obtido usando a estratégia Controle (quando as assinaturas receberam o mesmo peso), e que as atribuições de pesos maiores que 60% para a AEI e 90% para a ASD resultam em resultados significativamente inferiores aos resultados obtidos com a estratégia Controle. Em outras palavras, para cenários nos quais o número de atributos e correspondências é equilibrado, o melhor resultado foi alcançado ao se atribuir pesos iguais às assinaturas.

Por sua vez, os resultados para os cenários Medicamentos e Servidores Públicos mostram que a combinação de 80% para a assinatura ASD e 20% para a AEI resultaram em um resultado de qualidade superior ao da estratégia Controle. Este resultado pode ser interpretado da seguinte forma: em cenários com muitos atributos e nos quais existe um desbalanceamento entre o número de atributos das bases de dados (Razão dos atributos), a assinatura que considera os valores que o atributo assume (ASD) deve ser priorizada, recebendo um peso maior. Desse modo, a atribuição de pesos diferentes para as assinaturas proporcionou uma melhoria (pequena) nos resultados do PAC para estes cenários.

No entanto, um fato é comum aos dois grupos de resultados: a utilização de uma única assinatura, para todos os casos avaliados, sempre apresentou um resultado de qualidade inferior ao da utilização das duas assinaturas. Esta afirmação é ilustrada na Figura 5.6 na qual a primeira barra ($w_{aei} = 1$, $w_{asd} = 0$) e a última barra ($w_{aei} = 0$, $w_{asd} = 1$) não superaram a estratégia Controle para nenhum cenário.

Em resumo, os resultados apresentados nesta seção dão suporte às hipóteses H_5 (as duas assinaturas são relevantes para a abordagem) e H_6 (a atribuição de pesos distintos para as assinaturas gera resultados significativamente diferentes). Diante dos resultados alcançados pelo PAC foi constado que a abordagem proposta tem resultados melhores quando aplicada em cenários com bases de dados que apresentam um equilíbrio entre o número de atributos (total e correspondências).

5.6 Avaliação da Privacidade

Nesta seção é avaliada a privacidade da abordagem PAC considerando as suas limitações, o modelo de adversário simples (Honesto, porém curioso) e a utilização de um protocolo que

utiliza uma terceira parte confiável.

5.6.1 Discussão

Para avaliar a privacidade do PAC foi utilizado o paradigma da simulação [70] no contexto do PAC. Este paradigma avalia a privacidade da abordagem por meio das mensagens (informações) trocadas entre os participantes. Contudo, para que este paradigma seja utilizado é necessário que a simulação obedeça à seguinte diretriz [42]: as mensagens consideradas na simulação devem ser as mesmas que um adversário teria acesso e/ou utilizaria em um ataque real.

Desse modo, no contexto do PAC, os ataques perpetrados por adversários podem ser simulados em dois cenários: i) os participantes (excluindo a terceira parte confiável) utilizam as informações recebidas para quebrar a privacidade dos dados, ou ii) a terceira parte confiável utiliza as informações das assinaturas recebidas para quebrar a privacidade dos dados. Ressaltando que o PAC assume um modelo de adversário honesto porém curioso, modelo este que impede a terceira parte confiável de compartilhar informações (realizar conluio) com os demais participantes.

Para simular os ataques do primeiro cenário (ataques perpetrados pelos participantes) é necessário considerar que os parâmetros de anonimização são as únicas mensagens (informações) trocadas entre os participantes e que as assinaturas dos atributos são enviadas para a terceira parte confiável que, ao final da execução do PAC, notifica cada participante sobre quais atributos de sua base de dados estão presentes nas demais bases de dados. Desse modo, um adversário, de posse dos parâmetros de anonimização e da lista de atributos que têm em comum com os demais participantes, não é capaz de inferir qualquer informação extra dos demais participantes, além do que foi compartilhado pela terceira parte confiável. Portanto, é possível afirmar que a abordagem proposta é segura, ou seja, preserva a privacidade dos dados e do esquema contra os ataques perpetrados pelos participantes.

Por sua vez, para o último cenário, no qual a terceira parte confiável utiliza as informações das assinaturas para tentar quebrar a privacidade dos dados e do esquema dos demais participantes, deve ser considerado que a terceira parte confiável:

i não tem acesso às bases de dados dos demais participantes;

ii recebe as assinaturas com privacidade diferencial dos participantes.

Desse modo, por não ter acesso às bases de dados, a terceira parte confiável só dispõe das informações presentes nas assinaturas que, conforme comprovado no trabalho de Yean et al [72] e discutido por Lindell [42], garantem que as informações representadas têm privacidade diferencial, ou seja, não é possível identificar as entidades representadas pelas assinaturas. Portanto, em razão das assinaturas assegurarem a privacidade dos dados utilizados, a terceira parte confiável não é capaz de reidentificar os valores, a cardinalidade e o tipo do dado (numérico, textual, data, entre outros) dos atributos. Dessa forma, é possível afirmar que a abordagem proposta neste trabalho preserva a privacidade dos dados e do esquema em possíveis ataques realizados pela terceira parte confiável.

Em resumo, ao realizar a simulação dos ataques para os dois cenários foi confirmado que o PAC impede que os participantes e a terceira parte confiável aprendam ou infiram alguma informação, além do que foi repassado, sobre os dados dos demais participantes. Em outras palavras, o PAC é capaz de preservar a privacidade dos dados e do esquema. Desse modo, combinando os resultados desta avaliação e dos experimentos de qualidade conclui-se que é possível identificar/parear atributos em bases de dados distintas preservando a privacidade dos dados e do esquema, ou seja, a hipótese geral deste trabalho foi confirmada.

5.7 Avaliação da Eficiência

Nesta seção é apresentada uma investigação empírica sobre a eficiência do PAC. O objetivo principal deste conjunto de experimentos é avaliar o custo computacional da utilização do PAC em uma tarefa de REPP.

5.7.1 Desenho Experimental

O desenho experimental deste experimento consiste em executar o PAC com a estratégia de amostragem Conjunta com 2% do tamanho total de entidades das bases de dados, e medir o tempo de execução para cada cenário. Os resultados alcançados usando o PAC foram comparados com os resultados do DUMAS. Este último não considera a privacidade dos dados e utiliza todas as entidades das bases de dados.

5.7.2 Resultados

Os tempos de execução das abordagens PAC e DUMAS são ilustrados na Figura 5.7 para os diferentes cenários. O tempo de execução do DUMAS é representado por uma linha pontilhada vermelha, e seus valores são plotados no eixo vertical secundário, o eixo situado na margem direita da figura. Por sua vez, o tempo de execução do PAC é ilustrado por barras verticais, onde cada etapa do PAC é ilustrada por uma cor distinta.

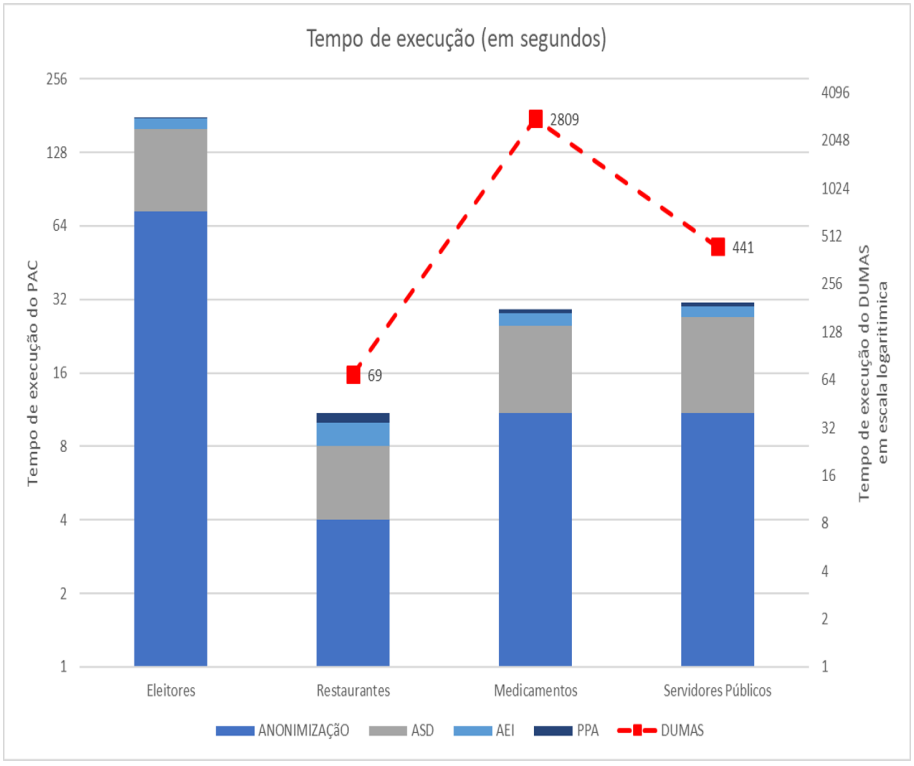


Figura 5.7: Tempo de execução das abordagens PAC e DUMAS.

5.7.3 Discussão

Os resultados expostos na Figura 5.7, ilustram que o tempo de execução do DUMAS foi superior ao do PAC para todos os cenários. É importante ressaltar que, para o cenário Eleitores, após quase 1 hora de execução, o DUMAS consumiu toda a memória disponível no computador e abortou a execução, ou seja, não foi capaz de identificar os atributos similares (correspondências). O tempo de execução excessivo do DUMAS pode ser explicado pelo fato do algoritmo utilizar os dados de todas as entidades para construir uma Matriz de Similaridades, a qual tem seu número de linhas e colunas definidos pelo número de entidades,

sendo esta a razão para o DUMAS necessitar de utilizar 112 GB de RAM para o cenário Eleitores.

Ao observar o tempo de execução do PAC, percebe-se que as etapas de Anonimização de dados, Criação de Assinaturas (ASD e AEI) e do Paramento Privado de Atributos (PPA) são responsáveis por consumir aproximadamente 33%, 66% e 1% do tempo total de execução, respectivamente. A Figura 5.8 ilustra o tempo de execução do PAC sem a etapa de de Anonimização de dados, destacando o tempo de execução das etapas de Criação de Assinaturas e PPA. É importante ressaltar que, para esta figura, os tempos de execução das etapas foram arredondados, o que fez com que o tempo de execução do PPA fosse o mesmo (0,1 segundo) para todos os cenários.

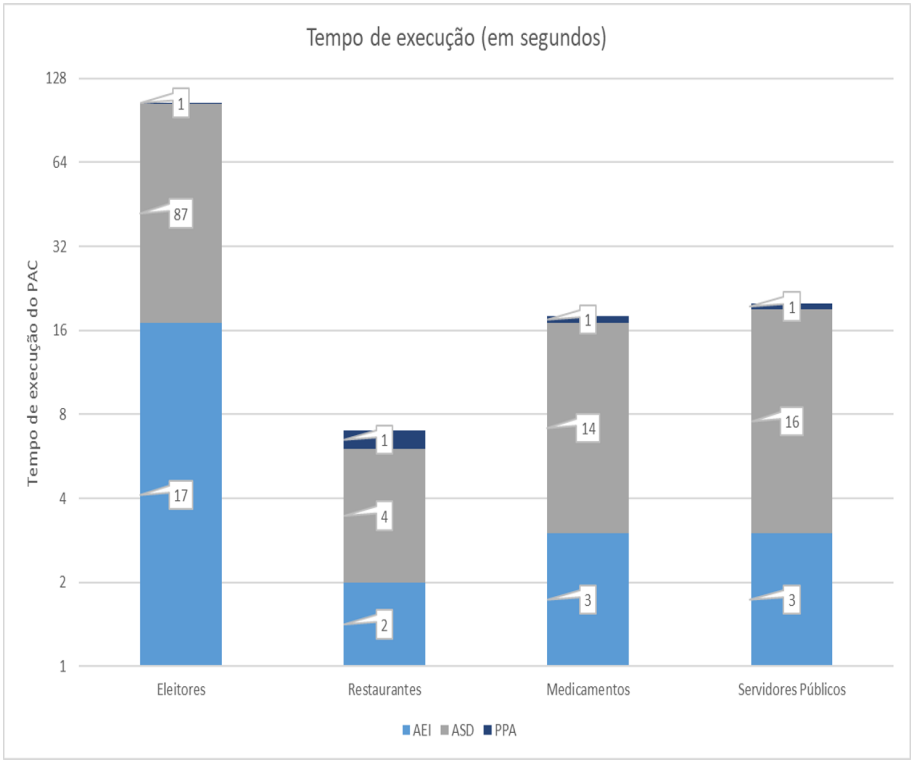


Figura 5.8: Tempo de execução do PAC, excluindo a etapa de Anonimização dos dados.

Os resultados ilustrados na Figura 5.8 mostram que a criação da assinatura ASD representa o maior custo computacional (considerando o tempo de execução) seguido pela criação da assinatura AEI e pela etapa de PPA. A diferença de tempos de execução da ASD em relação a AEI é explicado pela permutação (k) realizada pelo algoritmo de MinHash para gerar a assinatura ASD. O tempo de execução do PPA é constante e sofre influência do número de atributos das entidades. Esta influência pode ser observada para o cenário

Servidores Públicos (com 31 atributos), o qual apresentou o maior tempo de execução.

Desse modo, os resultados apresentados nesta seção mostram que o PAC é capaz de identificar os atributos similares (correspondências), considerando a privacidade durante todo o processo de maneira eficiente. Para o pior caso, bases de dados com um total de 15 milhões de entidades, o PAC foi capaz de identificar os atributos em menos de 3 minutos e, para os demais casos, o PAC realizou a identificação em menos de 30 segundos. Quando os resultados do PAC são comparados com um competidor (DUMAS), o qual não considera a privacidade dos dados, o tempo de execução do PAC é inferior para todos os cenários.

5.8 Considerações Finais

Os resultados dos experimentos e avaliações deste capítulo confirmaram todas as hipóteses experimentais ($H_1, H_2, H_3, H_4, H_5 e H_6$). Desse modo, com base nos resultados, é possível afirmar que o PAC é capaz de identificar atributos similares (correspondências), em bases de dados distintas, preservando a privacidade dos dados e dos esquemas. No entanto, os experimentos mostram que a eficácia do PAC é reduzida à medida que se aumenta o número de atributos não correspondentes nas bases de dados, ou seja, com bases nos experimentos o PAC deve ser utilizado para bases de dados com um número de atributos equivalentes.

Capítulo 6

Conclusões e Trabalhos Futuros

Neste capítulo são apresentadas as conclusões do trabalho e as principais perspectivas de trabalhos futuros.

6.1 Conclusões

O principal objetivo deste trabalho consistiu em propor uma abordagem (semiautomática), que adiciona uma nova etapa a REPP, capaz de realizar o pareamento de atributos (identificar os QIDs) preservando a privacidade dos dados. Desse modo, a hipótese geral de pesquisa do trabalho consistiu em:

H_{geral} : *É possível parear atributos, em bases de dados distintas, preservando a privacidade dos dados e dos esquemas em um contexto de REPP?*

Para investigar a hipótese geral do trabalho foram realizadas avaliações da qualidade (eficácia), eficiência e privacidade da abordagem proposta. Com base nos resultados das avaliações conclui-se que a abordagem proposta é capaz de:

i Parear atributos, de bases de dados distintas, com uma qualidade equivalente a uma abordagem que não considera a privacidade dos dados;

ii Preservar a privacidade dos dados e dos esquemas;

iii Realizar o pareamento de atributos de maneira eficiente, sendo capaz de realizar o pareamento de atributos para grandes bases de dados.

Diante dos resultados das avaliações a hipótese geral do trabalho (H_{geral}) foi confirmada, ou seja, a abordagem proposta neste trabalho foi capaz de parear os atributos, preservando a privacidade dos dados e dos esquemas. Os resultados da avalição também mostraram que a abordagem proposta apresentou uma queda nos resultados de qualidade para os cenários que utilizaram bases de dados com um desbalanceamento no número de atributos, ou seja, bases de dados com mais de 25 atributos não correspondentes. Mesmo com a hipótese geral de pesquisa confirmada, ainda existem melhorias a serem realizadas em trabalhos futuros, descritas com mais detalhes na próxima seção.

6.2 Trabalhos Futuros

Nesta seção, são elencadas quatro perspectivas de extensão do trabalho desenvolvido nesta dissertação, sendo elas;

Utilização de outros modelos de adversários. Para que o pareamento privado de atributos seja utilizado em tarefas de REPP que considerem outros modelos de adversários se faz necessário o aperfeiçoamento do PAC como, por exemplo, o modelo dissimulado com auditoria e malicioso. Nesse caso, a utilização de técnicas que permitam que os passos executados do PAC sejam auditados, e a utilização de algoritmos que propiciem uma maior segurança às assinaturas (a citar, Criptografia Homomórfica e *Secure Multi Party Computation*).

Pareamento de atributos compostos. Uma possível melhoria para este trabalho seria parear atributos compostos em diferentes bases de dados (1:n ou n:m). Por exemplo, uma base de dados representa os usuários do sistema com dois atributos (nome e sobrenome) e, em outra base de dados, os usuários são representados por apenas um atributo (nome completo). Nesse cenário a abordagem deveria identificar que nome completo é equivalente à combinação nome e sobrenome (nome completo = nome + sobrenome). Para tal, as assinaturas devem ser combinadas e o algoritmo de Pareamento Privado de Atributos (PAP) deve ser alterado.

Proposição de um algoritmo de amostragem específico para o pareamento de atributos. Este trabalho propôs uma metodologia para calcular o número de elementos das

amostras, contudo não foi realizada nenhuma investigação sobre como estes elementos são escolhidos (nas avaliações foi utilizada uma amostragem randômica). Assim, com base nos resultados dos experimentos que avaliaram o número de elementos das amostras, foi identificado que uma metodologia (algoritmo) própria de amostragem pode melhorar tanto a qualidade do pareamento como a privacidade dos dados. Nesse sentido, foi identificado como uma perspectiva de extensão desse trabalho o desenvolvimento de uma metodologia de amostragem que considere a privacidade dos elementos da amostra e tenha como objetivo maximizar a qualidade do pareamento.

Proposição de uma abordagem para escolha das chaves utilizadas na etapa de Blocagem da REPP. Um problema em aberto para a REPP é identificar como as entidades devem ser agrupadas para reduzir o custo computacional da tarefa, ou seja, como realizar a etapa de Blocagem de entidades. Um importante aspecto para realizar a blocagem é escolher o atributo (ou conjunto de atributos) que serão utilizados como chave para agrupar as entidades. Nesse sentido, a utilização das assinaturas de dados apresentadas nesse trabalho pode auxiliar na escolha dessa chave, utilizando a Entropia (AEI) para indicar qual atributo deve ser utilizado como chave de bloco. Desse modo, o desenvolvimento de uma abordagem para escolha das chaves utilizadas na etapa de Blocagem é listado como um trabalho futuro desta dissertação.

Bibliografia

[1] Babak Ahmadi, Marios Hadjieleftheriou, Thomas Seidl, Divesh Srivastava, and Suresh Venkatasubramanian. Type-based categorization of relational attributes. In *Proceedings of EDBT '09*, page 84, New York, New York, USA, 2009. ACM Press.

[2] Mohammad Alaggan, Sébastien Gambs, and Anne Marie Kermarrec. BLIP: Non-interactive differentially-private similarity computation on bloom filters. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7596 LNCS:202–216, 2012.

[3] Salgado Ana Carolina and Lóscio Bernadette Farias. Integração de Dados na Web. 2001.

[4] M R Anderlik and M a Rothstein. Privacy and confidentiality of genetic information: what rules for the new science? *Annual review of genomics and human genetics*, 2:401–433, 2001.

[5] Tiago Brasileiro Araujo, Carlos Eduardo Santos Pires, and Thiago Pereira da Nobrega. Spark-based Streamlined Metablocking. In *2017 IEEE Symposium on Computers and Communications (ISCC)*, pages 844–850. IEEE, 7 2017.

[6] Carlo Batini and Monica Scannapieco. *Data and Information Quality*. Data-Centric Systems and Applications. Springer International Publishing, 1 edition, 2016.

[7] Guilherme Dal Bianco, Renata Galante, Marcos Andre Goncalves, Sergio Canuto, and Carlos A. Heuser. A Practical and Effective Sampling Selection Strategy for Large Scale Deduplication. *IEEE Transactions on Knowledge and Data Engineering*, 27(9):2305–2319, 2015.

[8] Alexander Bilke and Felix Naumann. Schema Matching Using Duplicates. In *21st International Conference on Data Engineering (ICDE'05)*, pages 69–80. IEEE, 2005.

[9] Burton H. Bloom. Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM*, 13(7):422–426, 7 1970.

[10] BRASIL. *Constituição da Repíblica Federativa do Brasil*. Brasilia,DF, 1988.

[11] A.Z. Broder. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*, pages 21–29. IEEE Comput. Soc, 1997.

[12] Canada Vigilance Program. Canada Vigilance adverse reaction online database.

[13] Peter Christen. *Data Matching*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[14] Peter Christen. Data Linkage : Introduction , Recent Advances , and Privacy. (July), 2016.

[15] Peter Christen, Rainer Schnell, Dinusha Vatsalan, and Thilina Ranbaduge. *Efficient Cryptanalysis of Bloom Filters for Privacy-Preserving Record Linkage Peter*, volume 10235 of *Lecture Notes in Computer Science*. Springer International Publishing, Cham, 2017.

[16] Comitê Gestor da Infraestrutura Nacional de Dados Abertos. Portal Brasileiro de Dados Abertos.

[17] Conselho Federal de Médicina. Resolução 2.003/2012, 2012.

[18] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. 2006.

[19] Peter Diamond. Unemployment, Vacancies, Wages. *American Economic Review*, 101(4):1045–1072, 6 2011.

[20] Elizabeth A. Durham, Murat Kantarcioglu, Yuan Xue, Csaba Toth, Mehmet Kuzu, and Bradley Malin. Composite bloom filters for secure record linkage. *IEEE Transactions on Knowledge and Data Engineering*, 26(12):2956–2968, 2014.

[21] Elizabeth Ashley Durham. *A framework for accurate, efficient private record linkage*. PhD thesis, Vanderbilt University, 2012.

[22] European Commission. *Protection of personal data*. 2016.

[23] Niels Ferguson, Bruce Schneier, and Tadayoshi Kohno. *Cryptography Engineering Design Principles and Practical Applications*. Wiley, 2010.

[24] Martin Franke, Ziad Sehili, and Erhard Rahm. Parallel Privacy-Preserving Record Linkage using LSH-based blocking. In *International Conference on Internet of Things, Big Data and Security (IoTBDS)*, 2018.

[25] Srivatsava Ranjit Ganta, Shiva Prasad SP Kasiviswanathan, and Adam Smith. Composition attacks and auxiliary information in data privacy. *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '08*, pages 265–274, 2008.

[26] Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2 2009.

[27] Inc. Google. Google Flu Trends.

[28] Interagency Advisory Panel on Research Ethics Government of Canada. *Interagency Advisory Panel on Research Ethics*.

[29] IBGE. Nomes no Brasil, 2010.

[30] Anuj Jaiswal, David J. Miller, and Prasenjit Mitra. Schema matching and embedded value mapping for databases with opaque column names and mixed continuous and discrete-valued data fields. *ACM Transactions on Database Systems*, 38(1):1–34, 2013.

[31] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*, volume 103 of *Springer Texts in Statistics*. Springer New York, New York, NY, 2013.

[32] Wei Jiang, Chris Clifton, and Murat Kantarcioglu. Transforming semi-honest protocols to ensure accountability. *Data and Knowledge Engineering*, 65(1):57–74, 2008.

[33] Hasan KADHEM, Toshiyuki AMAGASA, and Hiroyuki KITAGAWA. MV-OPES: Multivalued-Order Preserving Encryption Scheme: A Novel Scheme for Encrypting Integer Value to Many Different Values. *IEICE Transactions on Information and Systems*, E93-D(9):2520–2533, 2010.

[34] Alexandros Karakasidis and Georgia Koloniari. Scalable Blocking for Privacy Preserving Record Linkage. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 527–536, 2015.

[35] Alexandros Karakasidis and Vassilios S. Verykios. A Sorted Neighborhood Approach to Multidimensional Privacy Preserving Blocking. In *2012 IEEE 12th International Conference on Data Mining Workshops*, pages 937–944. IEEE, 12 2012.

[36] Alexandros Karakasidis, Vassilios S. Verykios, and Peter Christen. Fake Injection Strategies for Private Phonetic Matching. pages 9–24. 2012.

[37] Dimitrios Karapiperis and Vassilios S. Verykios. An LSH-Based Blocking Approach with a Homomorphic Matching Technique for Privacy-Preserving Record Linkage. *IEEE Transactions on Knowledge and Data Engineering*, 27(4):909–921, 2015.

[38] Dimitrios Karapiperis, Vassilios S. Verykios, Eleftheria Katsiri, and Alex Delis. A Tutorial on Blocking Methods for Privacy-Preserving Record Linkage. pages 3–15. 2016.

[39] Hye-chung Kum, Ashok Krishnamurthy, Ashwin Machanavajjhala, and Stanley C Ahalt. Social Genome : Putting Big Data to Work. pages 56–63, 2014.

[40] Mehmet Kuzu, Murat Kantarcioglu, Elizabeth Durham, and Bradley Malin. A Constraint Satisfaction Cryptanalysis of Bloom Filters in Private Record Linkage. *Privacy Enhancing Technologies*, 6794:226–245, 2011.

[41] Mehmet Kuzu, Murat Kantarcioglu, Ali Inan, Elisa Bertino, Elizabeth Durham, and Bradley Malin. Efficient privacy-aware record integration. *Proceedings of the 16th International Conference on Extending Database Technology - EDBT '13*, page 167, 2013.

[42] Yehuda Lindell. *Tutorials on the Foundations of Cryptography*. Information Security and Cryptography. Springer International Publishing, Cham, 2017.

[43] Xiaoping Liu, Xiao-Bai Li, Luvai Motiwalla, Wenjun Li, Hua Zheng, and Patricia D Franklin. Preserving Patient Privacy When Sharing Same-Disease Data. *Journal of Data and Information Quality*, 7(4):1–14, 10 2016.

[44] D. G. Mayer and D. G. Butler. Statistical validation. *Ecological Modelling*, 68(1-2):21–32, 1993.

[45] Demetrio Gomes Mestre. *Uma Abordagem para Aprimoramento do Balanceamento de Carga do Método de Resolução de Entidades Standard Blocking baseado em MapReduce*. PhD thesis, Universidade Federal de Campina Grande, 2013.

[46] Noman Mohammed, Benjamin C M Fung, and Mourad Debbabi. Anonymity meets game theory: Secure data integration with malicious participants. *VLDB Journal*, 20(4):567–588, 2011.

[47] Alvaro E Monge and Charles P Elkan. The Field Matching Problem: Algorithms and Applications. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pages 267–270. AAAI Press, 1996.

[48] Milton Monti. PL 4060/2012 - Projetos de lei e sobre o tratamento de dados pessoais., 2012.

[49] Dale T Mortensen. Markets with Search Friction and the DMP Model. *American Economic Review*, 101(4):1073–1091, 6 2011.

[50] Frank Niedermeyer, Simone Steinmetzer, Martin Kroll, Rainer Schnell, Mehmet Kuzu, Murat Kantarcioglu, Elizabeth Durham, and Bradley Malin. Cryptanalysis of Basic Bloom Filters Used for Privacy Preserving Record Linkage. *Journal of Privacy and Confidentiality*, 6(2):59–79, 2014.

[51] Nobel Media AB. The Prize in Economic Sciences 2010 - Press Release, 2010.

[52] Thiago Pereira da Nóbrega, Carlos Eduardo Santos Pires, and Tiago Brasileiro Araujo. Avaliação Empírica de Técnicas de Comparação Privada Aplicadas na Resolução de

Entidades. In *Proceedings of the 31 st of the Brazilian Symposium on Databases (SBBD16)*, pages 121–126, 2016.

[53] North Carolina State Board of Elections. North Carolina Voters Registration.

[54] Ohio Secretary of State. Ohio Voters Registration.

[55] Payal Parmar, Shraddha B. Padhar, Shafika N. Patel, Niyatee I. Bhatt, and Rutvij H. Jhaveri. Survey of Various Homomorphic Encryption algorithms and Schemes. *International Journal of Computer Applications*, 91(8):26–32, 2014.

[56] Christopher A Pissarides. Equilibrium in the Labor Market with Search Frictions. *American Economic Review*, 101(4):1092–1105, 6 2011.

[57] Robespierre Pita, Clicia Pinto, Pedro Melo, Malu Silva, Marcos Barreto, and Davide Rasella. A Spark-based workflow for probabilistic record linkage of healthcare data. *CEUR Workshop Proceedings*, 1330:17–26, 2015.

[58] Huiling Qian, Jiguo Li, and Yichen Zhang. Privacy-Preserving Decentralized Ciphertext-Policy Attribute-Based Encryption with Fully Hidden Access Structure. pages 363–372. 2013.

[59] Shuo Qiu, Boyang Wang, Ming Li, Jesse Victors, Jiqiang Liu, Yanfeng Shi, and Wei Wang. Fast, Private and Verifiable: Server-aided Approximate Similarity Computation over Large-Scale Datasets. In *Proceedings of the 4th ACM International Workshop on Security in Cloud Computing - SCC '16*, pages 29–36, New York, New York, USA, 2016. ACM Press.

[60] Thilina Ranbaduge, Peter Christen, and Dinusha Vatsalan. Tree Based Scalable Indexing for Multi-Party Privacy-Preserving Record Linkage. *Australasian Data Mining Conference*, 2014.

[61] Thilina Ranbaduge, Dinusha Vatsalan, Peter Christen, and Vassilios Verykios. Hashing-Based Distributed Multi-party Blocking for Privacy-Preserving Record Linkage. volume 7301, pages 415–427. 2016.

[62] Sean M. Randall, Anna M. Ferrante, James H. Boyd, Jacqueline K. Bauer, and James B. Semmens. Privacy-preserving record linkage on large real world datasets. *Journal of Biomedical Informatics*, 50:205–212, 2014.

[63] Monica Scannapieco, Ilya Figotin, Elisa Bertino, and Ahmed K Elmagarmid. Privacy preserving schema and data matching. *Proceedings of the 2007 ACM SIGMOD international conference on Management of data (SIGMOD '07)*, pages 653–664, 2007.

[64] Rainer Schnell, Tobias Bachteler, and Jörg Reiher. Privacy-preserving record linkage using Bloom filters. *BMC Medical Informatics and Decision Making*, 9(1):41, 12 2009.

[65] LATANYA SWEENEY. k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 10 2002.

[66] US Department of Health & Human Service. *Health Information Privacy*. 2012.

[67] US Food and Drug Administration. FDA Adverse Event Reporting System.

[68] Dinusha Vatsalan and Peter Christen. Privacy-preserving matching of similar patients. *Journal of Biomedical Informatics*, 59(December):285–298, 2016.

[69] Dinusha Vatsalan, Peter Christen, Christine M O'Keefe, and Vassilios S Verykios. An Evaluation Framework for Privacy-Preserving Record Linkage. *Journal of Privacy and Confidentiality*, 6(1):75, 2014.

[70] Dinusha Vatsalan, Peter Christen, and Vassilios S. Verykios. A taxonomy of privacy-preserving record linkage techniques. *Information Systems*, 38(6):946–969, 2013.

[71] Dinusha Vatsalan, Ziad Sehili, Peter Christen, and Erhard Rahm. Privacy-Preserving Record Linkage for Big Data : Current Approaches and Research Challenges. In *Big Data Handbook*. Springer, 2016.

[72] Ziqi Yan, Jiqiang Liu, Gang Li, Zhen Han, and Shuo Qiu. PrivMin: Differentially Private MinHash for Jaccard Similarity Computation. 5 2017.

[73] Haina Ye, Xinzhou Cheng, Mingqiang Yuan, Lexi Xu, Jie Gao, and Chen Cheng. A survey of security and privacy in big data. In *2016 16th International Symposium on Communications and Information Technologies (ISCIT)*, pages 268–272. IEEE, 9 2016.

[74] Yelp. Yelp Dataset Challenge.

[75] Kassaye Yitbarek Yigzaw, Antonis Michalas, and Johan Gustav Bellika. Secure and scalable deduplication of horizontally partitioned health data for privacy-preserving distributed statistical computation. *BMC Medical Informatics and Decision Making*, 17(1):1, 2017.

Apêndice A

Detalhes do experimentos

A.1 Quadro com os parâmetros utilizados nos experimentos

Tabela A.1: Parâmetros dos experimentos

#	**Cenario**	Tamanho do Filtro de Bloom	Número de Permutações do MinHash
1	Eleitores	1024	128
2	Restaurantes	2048	128
3	Medicamentos	1024	128
4	Servidores Públicos	256	128

A.2 Como reproduzir os experimentos

Para reproduzir os experimentos primeiro acesse o GitHub (https://github.com/thiagonobrega/BAP) e siga as instruções disponíveis no Wiki do do GitHub.