CENTRO FEDERAL DE EDUCACÃO TECNOLÓGICA DE MINAS GERAIS CEFET-MG
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

ANONIMIZAÇÃO DE DADOS: UM COMPARATIVO DE ALGORITMOS DE PRIVACIDADE BASEADOS NO MODELO K-ANONYMITY

DIOGO SOUZA DE FIGUEIREDO

Orientador: Prof. Dr. Thiago de Souza Rodrigues

Belo Horizonte
Fevereiro de 2023

DIOGO SOUZA DE FIGUEIREDO

ANONIMIZAÇÃO DE DADOS: UM COMPARATIVO DE ALGORITMOS DE PRIVACIDADE BASEADOS NO MODELO K-ANONYMITY

Dissertação de mestrado submetido à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em Modelagem Matemática e Computacional do Centro Federal de Educação Tecnológica de Minas Gerais, como requisito final para obtenção do título de Mestre em Modelagem Matemática e Computacional.

Área de Concentração: Modelagem Matemática e Computacional

Linha de Pesquisa: Sistemas Inteligentes

Orientador: Prof. Dr. Thiago de Souza Rodrigues

Belo Horizonte
Fevereiro de 2023

Figueiredo, Diogo Souza de

F475a Anonimização de dados: um comparativo de algoritmos de privacidade baseados no modelo k-anonymity. / Diogo Souza de Figueiredo. – Belo Horizonte, 2023.
89 f. : il.

Dissertação (Mestrado) – Centro Federal de Educação Tecnológica de Minas Gerais, Programa de Pós-Graduação em Modelagem Matemática e Computacional, 2023.
Orientador: Prof. Dr. Thiago de Souza Rodrigues

1.Proteção de dados. 2. Algorítmos computacionais. 3. Anonimato. I. Rodrigues, Thiago de Souza. II. Centro Federal de Educação Tecnológica de Minas Gerais. III. Título

CDD 005.82

Elaboração da ficha catalográfica pela Bibliotecária Elisângela Gonçalves Barbieri CRB-6: 2624 / CEFET-MG

SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
COORDENAÇÃO DO PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

"ANONIMIZAÇÃO DE DADOS: UM COMPARATIVO DE ALGORITMOS DE PRIVACIDADE BASEADOS NO MODELO K-ANONYMITY"

Dissertação de Mestrado apresentada por **Diogo Souza de Figueiredo**, em 27 de fevereiro de 2023, ao Programa de Pós-Graduação em Modelagem Matemática e Computacional do CEFET-MG, e aprovada pela banca examinadora constituída pelos professores:

Prof. Dr. Thiago de Souza Rodrigues (Orientador)
Centro Federal de Educação Tecnológica de Minas Gerais

Prof. Dr. Sandro Renato Dias
Centro Federal de Educação Tecnológica de Minas Gerais

Prof. Dr. Gustavo Campos Menezes
Centro Federal de Educação Tecnológica de Minas Gerais

Visto e permitida a impressão,

Emitido em 27/02/2023

DECLARAÇÃO Nº 831/2023 - DECOM (11.56.03)

(Nº do Protocolo: NÃO PROTOCOLADO)

(Assinado digitalmente em 27/02/2023 18:09)
GUSTAVO CAMPOS MENEZES
PROFESSOR ENS BASICO TECN TECNOLOGICO
DCCN (11.58)
Matrícula: ###504#1

(Assinado digitalmente em 27/02/2023 17:10)
SANDRO RENATO DIAS
PROFESSOR ENS BASICO TECN TECNOLOGICO
DDE (11.48)
Matrícula: ###444#8

(Assinado digitalmente em 27/02/2023 16:00)
THIAGO DE SOUZA RODRIGUES
PROFESSOR DO MAGISTERIO SUPERIOR
DECOM (11.56.03)
Matrícula: ###518#3

Visualize o documento original em https://sig.cefetmg.br/documentos/ informando seu número: **831**, ano: **2023**, tipo: **DECLARAÇÃO**, data de emissão: **27/02/2023** e o código de verificação: **31c3e14e85**

AGRADECIMENTOS

Primeiramente a Deus por permitir trilhar esse caminho.

À minha esposa Rose, sempre companheira em todos os momentos, agradeço sempre, principalmente ao me incentivar, à realização deste trabalho.

Agradeço a toda minha família, Minha mãe, Meu Pai (in memoriam) e minha irmã, pessoas especiais em minha vida que contribuem e contribuíram, constantemente, no refinamento de meu caráter.

Agradeço também a todos os professores do CEFET-MG, em especial ao meu orientador Prof. Dr. Thiago de Souza Rodrigues pelos conselhos e por me orientar, com primor acadêmico, nessa caminhada.

Aos colegas e amigos, do CEFET-MG, em especial ao Denis Ribeiro e Wanderley Silva, pelas contribuições valorosas, pela ajuda nos estudos, trabalhos em grupo e por sempre ajudar a ter forças nos momentos de dificuldade. Agradeço as sugestões pertinentes para execução deste trabalho.

Ao CEFET-MG e a CAPES pelo apoio acadêmico e financeiro.

Ao meu grande amigo de trabalho Paulo Araújo, pela paciência e pelas orientações, muito obrigado por fazer sempre eu acreditar em mim mesmo quando eu me perdia em tantas demandas e nos momentos mais difíceis da minha vida.

Aos colegas e amigos do Centro Universitário de Belo Horizonte (UniBH), em Especial ao amigo Prof. Dr. Moisés Henrique Pereira Ramos pelo incentivo e contribuição para ingressar no Mestrado do CEFET-MG e aos demais amigos do mestrado, muito obrigado pela convivência e pelo apoio constante.

“O mundo não é um mar de rosas; é um lugar sujo, um lugar cruel, que não quer saber o quanto você é durão. Vai botar você de joelhos e você vai ficar de joelhos para sempre se você deixar. Você, eu, ninguém vai bater tão forte como a vida, mas não se trata de bater forte. Se trata de quanto você aguenta apanhar e seguir em frente, o quanto você é capaz de aguentar e continuar tentando. É assim que se consegue vencer. Agora se você sabe do teu valor, então vá atrás do que você merece, mas tem que estar preparado para apanhar. E nada de apontar dedos, dizer que você não consegue por causa dele ou dela, ou de quem quer que seja. Só covardes fazem isso e você não é covarde, você é melhor que isso.”

(Discurso do filme Rocky Balboa para o seu filho)

RESUMO

Em quase todas as atividades diárias, a internet está sempre presente. Muitas dessas atividades, requerem o compartilhamento de dados pessoais (por exemplo, endereço, telefone, idade, localidade, dentre outros) e dados considerados sensíveis (por exemplo, diagnóstico médico, conta de banco, número de documentos, dentre outros), gerando uma preocupação quanto a privacidade dessas informações perante a sociedade. Compartilhar informações exige a utilização de ferramentas de proteção de dados, de forma que as informações desses dados não possam ser utilizadas para identificar um indivíduo. Para atender as Leis e regulamentações aplicáveis, bem como, as políticas de proteção de dados informadas, surgiu um mecanismo de proteção de dados chamado anonimização de dados. Esse mecanismo consiste basicamente em remover os identificadores e ocultar dados sensíveis impossibilitando sua re-identificação. À medida que grandes quantidades de dados de indivíduos são disseminadas, novos desafios aparecem para a proteção de sua privacidade. Vários algoritmos de anonimização foram propostos, tornando-se a publicação de dados de preservação de privacidade em uma área de pesquisa bem abrangente. No entanto, é difícil identificar e selecionar o algoritmo mais adequado, devido ao grande número de algoritmos disponíveis e informações limitadas sobre seus desempenhos. Neste trabalho, são apresentados três algoritmos de anonimização chamados DataFly, Incognito e Mondrian. Será abordado suas eficiências no processamento dos dados, suas eficácias na quantidade de dados utilizados empregando um conjunto amplo de diferentes parâmetros, métricas e conjuntos de dados. Para orientar na seleção de um algoritmo, uma bateria de experimentos será realizada entre eles para identificar quais fatores podem influenciar no desempenho, apresentando as condições em que cada algoritmo supera uns aos outros em determinados requisitos de privacidade bem como suas vantagens e desvantagens.

PALAVRAS-CHAVE: anonimização de dados; k-anonymity; DataFly; Incógnito; Mondrian.

ABSTRACT

In almost every daily activity, the internet is always present. Many of these activities require the sharing of personal data (e.g. address, telephone, age, location, among others) and data considered sensitive (e.g. medical diagnosis, bank account, among others), generating a privacy concern. of this information to society. Sharing information requires the use of data protection tools, so the information in this data cannot be used to identify an individual. In order to comply with applicable laws and regulations, as well as informed data protection policies, a data protection mechanism called data anonymization has emerged. This mechanism basically consists of removing identifiers and hiding sensitive data, making it impossible to re-identify. As large amounts of data from individuals are disseminated, new challenges arise for the protection of their privacy. Various anonymization algorithms have been proposed, making publishing privacy-preserving data a very broad area of research. However, it is difficult to identify and select the most suitable algorithm, due to the large number of available algorithms and limited information about their performance. In this work, three anonymization algorithms called DataFly, Incognito and Mondrian are presented. Its efficiencies in data processing, its effectiveness in the amount of data used will be addressed, employing a wide set of different parameters, metrics and data sets. To guide the selection of an algorithm, a battery of experiments will be carried out among them to identify which factors can influence performance, presenting the conditions in which each algorithm outperforms each other in certain privacy requirements as well as their advantages and disadvantages.

KEYWORDS: data anonymization; k-anonymity; DataFly; Incognito; Mondrian.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANPD Autoridade Nacional de Preservação de Dados

EQ Classe de Equivalência (EC)

FREQc Frequência

IBGE Instituto Brasileiro de Geografia e Estatística

PPDP Publicação de dados de preservação de privacidade (*Privacy Preserving Data Publishing*)

LI Limite Inferior

LGPD Lei Geral de Proteção dos Dados

QID Quase Identificador

SA *Sensitive Atribute* (Atributo Sensitivo)

UI Limite Superior

VGH Hierarquia de Generalização de Valor (*Value Generalization Hierarchy*)

CAVG Métrica do Tamanho de Classe Equivalência (*Average Equivalence Class Size Metric*)

DM Métrica de Dicernibilidade (*Discernibility Metric*)

SUMÁRIO

Capítulo 1 INTRODUCÃO

Os volumes de dados gerados crescem exponencialmente a cada ano (GANTZ; REINSEL, 2020). Nesses dados, existem informações pessoais que crescem constantemente.

Com base em informações demográficas disponíveis, esse cenário tem chamado atenção de pessoas que visam criar serviços personalizados. Por essa razão, diversas empresas e organizações de vários setores, coletam dados pessoais para compartilhar de diferentes maneiras, sendo por questões legais ou até mesmo por razões monetárias. Isso trouxe novos desafios para as publicações dos conjuntos de dados com intuito de proteger a privacidade das pessoas.

Figura 1 - Adaptado de Visão geral da publicação de dados de preservação de privacidade.

Fonte:(AYALA-RIVERA et al., 2014)

Consequentemente, vários profissionais e inúmeros pesquisadores mostraram interesse na publicação de dados de preservação de privacidade (PPDP – Privacy Preserving Data Publishing). Na Figura 1, é apresentado um cenário típico de PPDP. Nela, são mostradas diversas fases do processamento dos dados. O modelo PPDP pressupõe que invasores podem descobrir informações confidenciais sobre os indivíduos entre os destinatários dos dados. Como consequência disso, o objetivo das técnicas do PPDP é reter a utilidade dos dados anônimos dos indivíduos modificando os dados, tornando-os menos específicos e

protegendo a privacidade dos indivíduos. O principal foco do PPDP é permitir a criação de conjuntos de dados que possam ter utilidades para tarefas distintas garantindo que no momento da publicação desses dados, eles sejam desconhecidos.

Por exemplo, no site Portal Brasileiro dos Dados Abertos (BRASIL, 2021), é praticamente impossível identificar quem são os destinatários dos dados publicados. Sendo assim, qualquer controlador de dados envolvido no compartilhamento de dados pessoais precisa aplicar os mecanismos de preservação da privacidade. Entretanto, esta não é uma tarefa trivial, uma vez que os profissionais não são necessariamente especialistas na área de privacidade de dados (GOODIN, 2014). Consequentemente, na maioria das vezes as organizações não possuem uma metodologia capaz de garantir o anonimato de maneira eficaz. Isso faz com que os profissionais apliquem métodos simples de desidentificação antes de liberar os dados para uso. Entre os métodos mais utilizados, podemos citar, por exemplo, a remoção de todos os identificadores diretos, como Nomes, CPF, RG, etc. Com isso, esta abordagem por si só, não é suficiente para preservar a privacidade dos dados (NARAYANAN; SHMATIKOV, 2008). Infelizmente, isso ocorre porque existe uma maneira de combinar conjuntos de dados distintos ou obter conhecimento prévio dos indivíduos, para realizar inferências sobre a origem de sua identidade. A maneira de se re-identificar um indivíduo é feito pela vinculação de atributos, que na maioria das vezes são conhecidos como quase identificadores (QIDs), como CPF, RG, sexo, etc. Ela pode ocorrer pela combinação de informações complementares. O uso de um segundo banco de dados com informações distintas, pode servir de base para fazer cruzamentos de informações. Um exemplo, seria a combinação do código postal e do modelo do veículo. Cruzamento de dados indiretos, mas não diretamente coligados e não referenciados podem ser utilizados como complemento para um procedimento de re-identificação. Esse é um exemplo das possibilidades do cruzamento das Informações auxiliares que podem abrir espaço para uma possível re-identificação.

Muitos casos de violação de privacidade acontecem com frequência e geram uma preocupação muito grande quanto aos custos e prejuízos financeiros que são causados. Além da questão financeira, as empresas e os usuários podem ter suas imagens e credibilidades prejudicadas, tendo em vista que as empresas deveriam garantir a privacidade dos usuários.

De acordo com o estudo realizado pelo Ponemon Institute, em 2018 foram verificados os principais motivos das causas de violações de privacidade e também os prejuízos gerados (PONEMON INSTITUTE, 2018). O estudo apresentou pesquisas realizadas em 419 organizações. Os países que participaram da pesquisa foram: Brasil, França, Canadá, Estados Unidos, Reino Unido, Arábia Saudita, Alemanha, Japão, Itália, Índia, Singapura, Indonésia, Emirados Árabes Unidos, Austrália, África do Sul, Filipinas e Malásia. Concluíram que 27% dos incidentes foram causados por erro humano, 25% foram relacionadas as falhas nos

sistemas, que englobam falhas nos processos de negócios e 48% relacionados aos ataques criminosos. Foram apontados também os prejuízos financeiros. O custo médio da violação de privacidade gira em torno de 3,86 milhões de dólares. Esse custo apresentou um aumento de 6,5% em relação ao ano anterior. Estima-se que o custo médio de registro perdido fique próximo a 148 dólares. Um dado relevante identificado pelo estudo, mostra que a economia média com os gastos na contratação de uma equipe para ação rápida em caso de incidentes é de 14 dólares por registro. Isso representa um custo médio de 10% por registro.

Com os dados apresentados, levando-se em conta a questão financeira, fica evidente que a prevenção é mais recomendada do que correr o risco de incidentes com vazamento de informações. Os incidentes podem apresentar um gasto muito maior quanto a pagamentos de valores provindos de processos, etc.

Outro estudo realizado por Data Breaches (QUICK et al., 2018), afirma que uma grande variedade de empresas, governos e universidades tiveram mais de 400 casos públicos de violação de dados que foram identificados entre os anos de 2004 e 2016. Alguns casos de violação de privacidade chamam atenção por conta da exposição dos dados, que se tivessem sido tratados e adotados uma boa estratégia de anonimização, poderiam ter sido evitados.

Em 2008, durante a semana de prevenção de fraudes de identidade nacional foi anunciado que um disco rígido contendo várias informações pessoais dos membros das forças armadas do Reino Unido, como por exemplo, dados bancários, passaporte, etc, desapareceu (BBC, 2008).

Em 2012, acidentalmente, um funcionário da imigração australiana enviou aos organizadores da Copa de Futebol Asiático os números de passaportes, números de identificação de todos os lideres, etc, incluindo os dados do presidente dos Estados Unidos da América, Barack Obama. O presidente compareceria à reunião do G20 em Brisbane (QUICK et al., 2018). No mesmo ano, um notebook roubado pode ter sido a razão da exposição de nomes, datas de nascimento, endereços, números de telefone, etc de clientes do sistema de saúde americano (MCCANN, 2013).

Em 2014, na cidade de New York, dados de aproximadamente 173 milhões de viagens de táxi ocorridas foram vazados devido à falta de tratamento correto no processo de anonimização das informações. Os dados que foram divulgados apresentaram os locais de partida e chegada de cada táxi, documentos de identificação dos motoristas, prefixo do veículo utilizado e algumas outras informações consideradas relevantes (PANDURANGAN, 2014).

Em 2015, a Agência Sueca de Transportes (Transportstyrelsen), ao terceirizar sua manutenção de Tecnologia da Informação, disponibilizou os dados de milhões de carteiras de motoristas para profissionais de Tecnologia da Informação. Esse caso chegou às manchetes

da época na Suécia quando a ex-diretora-geral da agência, foi multada devido à falta de proteção adequada das informações consideradas sigilosas (THE LOCAL, 2017).

Como resposta aos diversos problemas decorridos de vazamentos de dados, vários algoritmos de anonimização têm sido apresentados na área de PPDP. Contudo, a definição de um algoritmo ótimo para um possível cenário de publicação é desafiadora para os profissionais. Não existe somente uma infinidade de algoritmos de anonimato que podem ser utilizados, existe também a questão de reinvidicação de superioridade única de cada algoritmo novo lançado. As avaliações desses algoritmos são limitadas, pois se direcionam em apresentar os benefícios do algoritmo atual em relação a algoritmos propostos anteriormente. Na maioria das vezes, esse cenário prejudica o processo de avaliação quanto às configurações empregadas nos experimentos, como por exemplo, ao fazer um comparativo, utilizar somente um tipo de métrica e omitir outros parâmetros de verificação. Além disto, conforme novas métricas são adicionadas, essas novas métricas favorecem o algoritmo proposto quanto a seus aspetos particulares.

As situações citadas acima podem gerar dúvidas e até mesmo confundir os profissionais, induzindo-os à erros. Independente dos parâmetros utilizados na entrada, se um determinado algoritmo supera o outro em uma métrica escolhida, esse algoritmo tende a ser escolhido como o melhor para a situação envolvida. Em consequência, esse comportamento não é garantido. Um algoritmo pode ter seu desempenho afetado à medida que um novo conjunto de dados de entrada com características novas é anonimizado. Com todos os desafios presentes, existe uma grande necessidade de expandir as avaliações desses algoritmos visando estender um conjunto mais amplo de configurações experimentais.

1.1 Objetivos

O objetivo geral deste trabalho é comparar modelos de anonimização de dados (modelos de privacidade) de três algoritmos, Datafly, Incognito e Mondrian, utilizando as técnicas de generalização e supressão. Mostrar uma revisão abrangente dos diferentes modelos e técnicas de transformação de dados utilizando uma ferramenta prática existente que pode ser usada para aplicar essas técnicas para preservar a privacidade de um indivíduo e apresentar seu grau de eficiência. Acredita-se que, com isso, pesquisadores e profissionais podem se basear nos resultados que serão apresentados para escolher algumas das técnicas de anonimização apropriadas com base em vários aspectos, como tipo de dados, nível de privacidade, perda de informações e comportamento do algoritmo.

O objetivo específico é apresentar aos profissionais explicações detalhadas das variações de desempenho dos algoritmos apresentados. Apresentar o conceito de PPDP e demonstrar a aplicação de uma ferramenta prática.

Além da abordagem de anonimização de dados e revisão de literatura, pretende-se atingir os seguintes objetivos específicos:

- Caracterização dos algoritmos que serão utilizados.
- Seleção da base de dados a ser utilizada.
- Filtragem e preparação da base de dados.
- Definição dos parâmetros a serem utilizados na comparação.
- Execução dos testes.
- Comparação dos resultados.

1.2 Organização Trabalho

Este trabalho está organizado da seguinte forma: O Capítulo 2 apresenta a caracterização do problema, os conceitos básicos relacionados a anonimização que serão definidos e discutidos. O Capítulo 3 apresenta a metodologia de pesquisa e os principais métodos e técnicas de aplicação bem como a seleção dos algoritmos. O Capítulo 4 apresenta os experimentos realizados com os algoritmos. Os resultados obtidos serão apresentados e discutidos neste capítulo. O Capítulo 5 faz uma análise de resultados. Finalmente, o Capítulo 6 conclui este trabalho e apresenta trabalhos futuros.

Capítulo 2 CONTEXTUALIZAÇÃO

2.1 Caracterização do problema

A anonimização de dados é um processo que tem como objetivo remover ou mascarar os dados confidenciais de um indivíduo, preservando seu formato original. Este processo tem grande importância, pois ao compartilhar dados externamente, deve-se garantir que qualquer informação sensível, seja ele de origem de banco de dados ou de documentos confidenciais, não sejam expostas.

Mensagens, e-mail, artigos de jornais ou relatórios, são um tipo especial de documento onde os dados estão na forma não estruturada e são representados em forma de linguagem natural.

É necessário identificar no conteúdo desses documentos as estruturas do texto que representam nomes ou identificadores exclusivos, chamados de entidades. Existe uma preocupação ao compartilhar informações de terceiros sem promover a violação de privacidade.

2.2 Lei Geral de Proteção dos dados

LEI GERAL DE PROTEÇÃO DE DADOS - LGPD (BRASIL, 2018), é a lei nº 13.709, aprovada em agosto de 2018 pelo congresso nacional do Brasil e com vigência a partir de agosto de 2020. Tem como principal objetivo proteger os direitos fundamentais de liberdade e de privacidade e o livre desenvolvimento da personalidade da pessoa natural, além de focar na criação de um cenário de segurança jurídica padronizando regulamentos e visando promover a proteção aos dados pessoais de qualquer indivíduo que esteja no território nacional obedecendo parâmetros internacionais existentes.

A lei estabelece algumas regras quanto à coleta e manutenção das informações dos dados. Essa coleta deve sempre ser realizada com o consentimento do indivíduo, exceto em casos de mandados judiciais ou casos de investigações criminais no intuito de garantir a

segurança pública ou do Estado. Essas regras valem para dados digitais originados de internet ou outros meios.

Dados considerados sensíveis, como por exemplo, estado de saúde, religião, características físicas, entre outros, foram classificados como dados restritos. Esses dados, não podem ser utilizados para práticas que podem levar a situações de caráter discriminatório e obrigatoriamente devem ser protegidos. Obrigatoriamente, dados considerados médicos, só podem ser utilizados em casos onde existe uma autorização prévia do indivíduo, não podendo ser utilizados para práticas visando fins comerciais. Órgãos públicos e empresas privadas são obrigados a informar aos usuários, quanto ao tratamento dos dados e compartilhamentos com terceiros. Para a finalidade de compartilhamento, o indivíduo obrigatoriamente deve conceder uma autorização para que essa prática possa ser executada.

Empresas privadas ou Órgãos públicos, devem fornecer ao indivíduo, ferramentas que permitam o acesso aos dados, possibilitando correções, exclusão ou transferências de serviços, tendo como base o serviço de portabilidade. Pessoas físicas não serão afetadas caso os dados sejam utilizados para fins artísticos, acadêmicos, pessoais, desde que seja respeitado a condição anônima de algumas informações consideradas sensíveis, como por exemplo, endereço, CPF, etc. Empresas que trabalham com tratamento de dados, seja no próprio território nacional ou exterior, precisam se adequar a LGPD, independente se suas operações estão sendo realizadas dentro ou fora do país. Empresas como Facebook e Google, por exemplo, respondem à LGPD mesmo se fizerem coleta de dados no território brasileiro para realizar o tratamento desses dados fora do país. Ao enviar os dados para fora do país, empresas com sedes no exterior, podem realizar essas transferências desde que o país de destino, possua leis abrangentes quanto a questão de privacidade ou disponha de mecanismos similares de proteção baseados na legislação brasileira.

Toda empresa, desde que não seja por determinação da Lei ou razão que seja justificada, deve garantir que, caso os dados de algum indivíduo não sejam mais relevantes, esses dados devem ser deletados completamente da base de dados. Por exemplo, o encerramento de uma conta de rede social que foi fechada a pedido do usuário.

Empresas que tiverem vazamentos de dados, serão analisadas e julgadas pela Lei nº 13.853/2019 da AGÊNVIA NACIONAL DE PROTEÇÃO DE DADOS – ANPD - (BRASIL, 2019). De acordo com a gravidade do caso, as empresas serão obrigadas a informar as autoridades, as falhas ocorridas no momento em que foi identificado o vazamento dos dados. Após esse comunicado, não são mais permitidas manutenções antes de informar publicamente o motivo do vazamento. A divulgação pública ou não divulgação de vazamento, pode variar de acordo com a situação. As multas são aplicadas de acordo com a proporcionalidade de cada situação. Podem ser em caráter de advertência ou multas que podem variar de 2% sobre o faturamento

anual ou até um valor máximo de R$ 50.000.000,00. Multas diárias também podem ser aplicadas, desde que a soma do valor total não seja maior que o valor máximo de R$ 50.000.000,00.

A LGPD apresenta dois tipos de responsabilidades. A responsabilidade administrativa e responsabilidade civil. A responsabilidade administrativa prevê sanções dispostas pelo artigo 52, através da ANPD. Já a responsabilidade civil é baseada no artigo 42 da LGPD, atribuindo a responsabilidade de reparação por danos causados pelo uso de dados privados.

A ANPD, pode fiscalizar e aplicar penalidades quanto ao descumprimento da LGPD, ficando responsável por regular e orientar as empresas quanto a aplicação da Lei. A LGPD, exige o uso de agentes de tratamento de dados. Algumas funções são exigidas, como por exemplo, a função de controlador de dados; ele tem como tarefa principal, tomar as decisões sobre o tratamento a ser utilizado, função de operador; executar o tratamento de acordo com os parâmetros informados pelo controlador e o encarregado, que faz o contato com os titulares dos dados e com a autoridade nacional. Para o caso de falhas, o responsável pelos dados deve adotar normas de segurança, auditorias, identificar e corrigir vulnerabilidades com rapidez, apresentar planos de contingência e ter controle quanto ao contato à ANPD sobre indivíduos afetados e violações.

2.3 Anonimização de dados

Antes de apresentar o conceito de anonimização de dados, é preciso, com base na LGPD, definir o que é Dado Pessoal. O conceito, apesar de ser bastante abrangente, pode ser definido da seguinte forma: é a característica que identifique ou que permita identificar um indivíduo (por exemplo, nome, sobrenome, data de nascimento, documentos pessoais como CPF, RG, CNH, Carteira de Trabalho, passaporte e título de eleitor, endereço residencial ou comercial, telefone, e-mail, cookies e endereço IP, entre outros), bem como, fragmentos de informação.

A definição de Anonimização de Dados, pode ser definida com base no artigo 5º da LGPD:

“Anonimização: utilização de meios técnicos razoáveis e disponíveis no momento do tratamento, por meio dos quais um dado perde a possibilidade de associação, direta ou indireta, a um indivíduo”.

O processo de anonimização visa mascarar ou ofuscar os dados antes destes serem disponibilizados ou compartilhados, utilizando técnicas para que os indivíduos que tiveram os dados anonimizados não possam ser identificados novamente (BASSO; MATSUNAGA et al., 2016), ou seja, o processo de anonimização deve ser irreversível, não podendo o dado em hipótese alguma, ser revertido ou recuperado.

A LGPD não determina quais padrões e técnicas que podem ser usados em processos de anonimização. Para essa questão, a **ANPD** fica responsável pelo assunto.

2.4 Atributos para Anonimização

Antes de iniciar o processo de anonimização de um dado, é preciso definir quais atributos (dados) devem ser anonimizados e quais técnicas serão aplicadas a cada um deles. Os atributos necessitam passar por um processo de classificação de acordo com a sensibilidade que cada informação apresenta, para que o atributo possa ser divulgado ou compartilhado.

A classificação dos atributos está dividida em: (i) Atributos identificadores, que identificam os indivíduos (por exemplo, nome, CPF, RG); (ii) atributos semi-identificadores, que, se combinados com informações externas, expõem indivíduos ou aumentam a certeza sobre suas identidades (por exemplo, data de nascimento, CEP, cargo, tipo sanguíneo); e (iii) atributos sensíveis, que se referem a condições específicas dos indivíduos (por exemplo, salário, exames médicos, doença) (CAMENISCH; FISCHER-HÜBNER; RANNENBERG, 2011).

A Tabela 1, nos fornece as principais informações do conjunto de dados anonimizados:

- Registros (Tuplas – ID – FreqC (Frenquência)): é a quantidade de registros (linhas) existentes na tabela. As Tabelas 1 (A) e 1 (B) possui 4 registros cada uma.
- EQs ou ECs - Classes de Equivalência: Especifica a quantidade de agrupamentos de acordo com os atributos escolhidos: Por exemplo: Na Tabela 1 (B) a Tupla (ID) 1 e 4 formam o EQ 1, a Tupla (ID) 2 e 3 formam o EQ 2 com

os registros iguais nos atributos Nome, Idade e CEP.

- QIDs - Quase Identificadores: São os atributos escolhidos para o processo de anominização. Nas Tabelas 1 (A) e 1 (B) os atributos “Nome, Idade e CEP” foram escolhidos como quase identificadores e possuem uma configuração que assume o valor de QIDs = 3.
- VGH: Hierarquia de generalização de valor. São todas variações que cada atributo possui. Por exemplo: o atributo CEP possui um VGH com profundidade 2 (30.421-0** e 33.562-0***).

Tabela 1 - Exemplo de tabela anonimizada.

Tupla (id)	**Nome**	**Idade**	**CEP**	**DOENÇA**
1	Rodrigo	27	30.421-049	Sem doença
2	Renata	19	33.562-078	Infecção viral
3	Luciana	18	33.562-025	Infecção viral
4	Paulo	23	30.421-050	Tuberculose

(A) TABELA ORIGINAL

Tupla (id)	**Nome**	**Idade**	**CEP**	**DOENÇA**
1	*	20 < Idade ≤ 30	30.421-0**	Sem doença
4	*	20 < Idade ≤ 30	30.421-0**	Tuberculose
2	*	Idade ≤ 20	33.562-***	Infecção viral
3	*	Idade ≤ 20	33.562-***	Infecção viral

(B) TABELA ANONIMIZADA

Fonte: (O próprio autor)

2.5 Técnicas de Anonimização

Com base nos atributos identificados pela sua sensibilidade de divulgação (identificadores, semi-identificadores e sensíveis), as técnicas de anonimização podem ser aplicadas para proteger a identidade dos indivíduos (OHM, 2010). Abaixo será apresentado os conceitos de algumas técnicas mais conhecidas como: Supressão, Generalização e Agregação. Outras técnicas da literatura podem ser citadas também, como por exemplo, Criptografia, Mascaramento, Substituição, Embaralhamento e Anulação. (BRANCO JR; MACHADO; MONTEIRO, 2014).

2.5.1 Supressão

É a retirada completa de um atributo. Uma coluna inteira correspondente aos dados a serem anonimizados é totalmente excluída. Essa técnica é mais utilizada nos dados identificadores.

Na Tabela 2, os campos Nome, Nascimento, Gênero e Código Postal são atributos que podem ser totalmente excluídos.

Tabela 2 - Dados originais - Adaptado de Supressão de atributos.

Nome	Raça	Nascimento	Gênero	Código Postal	Reclamação
Daniel	Negro	02/14/1965	Masculino	02141	Dor no peito
Kate	Negro	10/21/1965	Feminino	02138	Olho dolorido
Maria	Negro	08/24/1965	Feminino	02138	Chiado no ouvido
Flavio	Branco	10/18/1964	Masculino	02138	Falta de ar
Felipe	Branco	08/13/1964	Masculino	02139	Articulação dolorida
Jairo	Branco	05/05/1964	Masculino	02139	Febre
Rodrigo	Branco	02/13/1967	Masculino	02138	Vômito
Adriano	Branco	03/21/1967	Masculino	02138	Dor nas costas

Fonte: (OHM, 2010)

Enquanto que na Tabela 3, os campos Raça e Reclamação são preservados no conjunto de dados.

Tabela 3 - Dados preservados - Adaptado de Supressão de atributos.

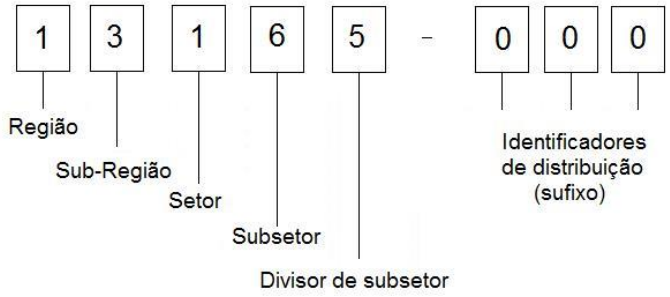
Raça	Reclamação
Negro	Dor no peito
Negro	Olho dolorido
Negro	Chiado no ouvido
Branco	Falta de ar
Branco	Articulação dolorida
Branco	Febre
Branco	Vômito
Branco	Dor nas costas

Fonte: (OHM, 2010)

2.5.2 Generalização

É uma ótima opção para buscar o equilíbrio entre utilidade e privacidade. Também chamada de recodificação, tem por objetivo substituir os valores de um atributo por valores menos específicos ou apenas manter parte dos dados, ao invés de excluí-los totalmente. Podemos citar o CEP, como exemplo, cujos números representam o escopo geográfico. De acordo com as regras, quanto a mais à esquerda for o número, maior seu alcance. A Figura 2 apresenta a forma como o CEP foi dividido no território brasileiro.

Figura 2 - Estrutura do CEP.



Fonte: (CORREIOS, 2020)

De acordo com as divisões de dígitos, fica a seguinte regra:

- primeiro dígito, mais à esquerda, representa a região geográfica.
- segundo dígito, representa a sub-região.
- terceiro dígito, representa o setor.
- quarto dígito, representa o subsetor.
- quinto dígito representa o divisor do subsetor.
- os últimos três números à direita, representam os identificadores de distribuição (CORREIOS, 2020).

O país foi dividido em dez regiões postais para fins de codificação, utilizando como parâmetro o desenvolvimento socioeconômico e fatores de crescimento demográfico de cada Unidade da Federação ou conjunto delas. De acordo com a figura 3, a distribuição do CEP foi feita no sentido anti-horário a partir do estado de São Paulo, pelo primeiro algarismo.

(CORREIOS, 2020).

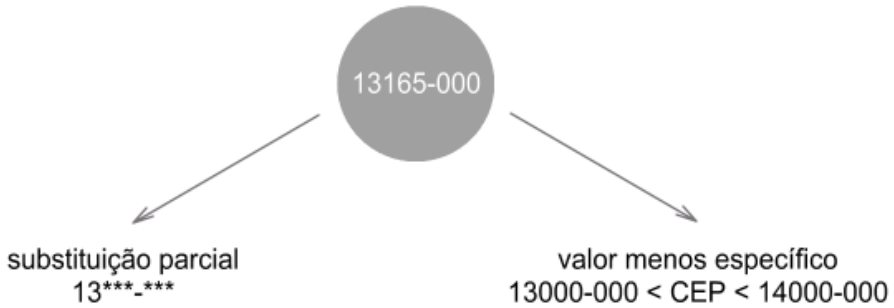
Figura 3 - Divisão das Regiões.

Fonte: (CORREIOS, 2020)

Como exemplo, podemos aplicar a generalização no CEP: 13165-000 utilizando a substituição parcial. Após a transformação, temos o seguinte resultado: CEP: 131**-***. Desta maneira, estamos aumentando a cobertura geográfica da região de Engenheiro Coelho / SP para o setor de Campinas e seu entorno.

Outra maneira de aplicar a generalização é substituir o atributo por valor menos específico. Um exemplo desse modelo é apresentado na Figura 4.

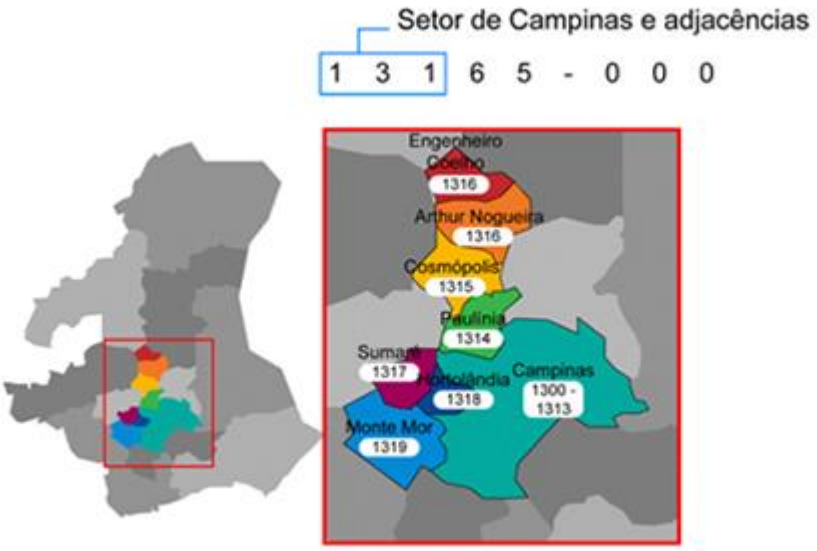
Figura 4 – Formatos de Generalização.



Fonte: (O próprio autor)

A Figura 5 mostra diferentes regiões identificadas, considerando apenas a parte mais à esquerda do CEP, pelos dígitos 131, reduzindo a possibilidade de re-identificação do endereço (OHM, 2010).

Figura 5 - CEP setor de Campinas (São Paulo/Brasil) e adjacências.



Fonte: (CORREIOS, 2020)

2.5.3 Agregação

É a ação de difundir dados estatísticos resumidos protegendo os indivíduos contra a re-identificação. Uma aplicação da técnica de agregação é mostrada na Tabela 4. Com base no valor anonimizado, foram selecionados 2 indivíduos da coluna Gênero que são masculinos e possuem Falta de ar.

Tabela 4 - Adaptado de Tabela anonimizada por agregação.

Nome	Raça	Nascimento	Gênero	Código Postal	Reclamação
Lucas	Negro	09/20/1965	Masculino	02141	Falta de ar
Daniel	Negro	02/14/1965	Masculino	02141	Dor no peito
Kate	Negro	10/21/1965	Feminino	02138	Olho dolorido
Helen	Negro	11/07/1964	Feminino	02138	Articulação dolorida
Flavio	Branco	10/18/1964	Masculino	02138	Falta de ar

Fonte: (OHM, 2010)

Para não correr o risco de identificação do indivíduo, ao realizar o processo de agregação, as consultas não podem retornar registros exclusivos. Nessas técnicas, pode-se utilizar médias, somas, que são comuns, possibilitando seu uso em bancos estatísticos.

Um bom exemplo, são os dados do censo demográfico realizado pelo IBGE (Instituto Brasileiro de Geografia Estatística). Relatórios consolidados são gerados mostrando o perfil socioeconômico das regiões brasileiras, sem identificar os indivíduos (IBGE, 2010).

2.5.4 Mascaramento

É a divulgação de dados com valores modificados. A anonimização de dados é feita implementando estratégias de alteração, como embaralhamento de caracteres, termo ou substituição de caractere. Um valor numérico pode ser substituído por um símbolo como “*” ou “x”. Isso dificulta a identificação ou engenharia reversa. Exemplo: alterar a data de nascimento 15/08/1985 para ** / ** / 19 **.

2.5.5 Substituição

Todos os dados são completamente substituídos por valores aleatórios de uma lista de dados falsos, mas de aparência semelhante. Por exemplo, os sobrenomes podem ser trocados por outros sobrenomes não relevantes ou os números de cartão de crédito podem ser substituídos por uma sequência aleatória de 16 números. Para aproveitar esse método de maneira adequada, os usuários devem ter listas igual ou superior à quantidade de dados que estão tentando tornar anônimos. Exemplo: número original do cartão: 1452 5689 6574 8596. Resultado após a substituição: 1052 5749 6524 8326

2.5.6 Embaralhamento ou Permutação

Essa técnica envolve a reorganização dos dados para que os atributos de dados permaneçam presentes, mas não correspondam a seus registros originais. Costuma ser equiparado ao embaralhamento de um baralho de cartas. Este método é eficaz quando não há necessidade de avaliar os dados com base nas relações entre as informações contidas em cada registro. Exemplo: CPF: 065.956.968-98. Com embaralhamento 898.696.595-60

2.6 K-anonymity

Um dos maiores riscos relacionados à anonimização por generalização, do ponto de vista individual, é deixar registros expostos à processos simples de re-identificação em razão do fato de ser o único com determinados atributos em uma base de dados (NARAYANAN; SHMATIKOV, 2010).

O modelo k-anonymity foi o primeiro modelo de anonimização proposto na literatura (SAMARATI; SWEENEY, 1998). Ele apresenta o formato que o conjunto de dados deve possuir após a anonimização com o intuito de impossibilitar a re-identificação de indivíduos que integram os conjuntos de dados. Para tanto, a informação de cada pessoa no conjunto de dados pós-anonimizados não pode ser distinguida em pelo menos k-1 indivíduos, cujas informações também aparecem nesse mesmo conjunto (SWEENEY, 2002).

Sua escolha se deve ao fato de que o modelo serve de base para propor modelos mais seguros e também tem o princípio fundamental de privacidade. Sua facilidade de aplicação o tornou-o amplamente discutido, pois possibilita que novas técnicas de anonimização possam ser aplicadas em diferentes contextos, como por exemplo, controle de divulgação estatística redes sociais, área da saúde, base de localização serviços, mineração de dados (IYENGAR, 2002), etc.

Por fim, os algoritmos de anonimização k selecionados para este trabalho, são amplamente utilizados como base pela comunidade e tornaram-se muito representativos.

Na Tabela 5, ao fazer uma busca para a Identificação de um indivíduo, percebe-se que não é possível reidentificar o atributo Nome de um indivíduo depois de um processo de k-anonymity. Após a generalização, perdeu-se o atributo que liga aos demais atributos, passando a ter mais possibilidade de correspondências. Apesar da eficiência, a técnica não está livre dos riscos de re-identificação.

Tabela 5 - Adaptado de Tabela anonimizada por k-anonymity.

	Quase-Identificadores		
Id	**Gênero**	**Ano Nascimento**	**Resultado do Teste**
1	Masculino	1950-1959	+ve
2	Masculino	1960-1969	-ve
4	Masculino	1950-1959	-ve
6	Feminino	1970-1979	-ve
7	Feminino	1960-1969	+ve
9	Masculino	1970-1979	-ve
10	Masculino	1970-1979	-ve
11	Feminino	1960-1969	-ve
12	Masculino	1950-1959	+ve
13	Feminino	1970-1979	-ve
14	Masculino	1960-1969	-ve

Fonte: (EL EMAM; DANKAR, 2008)

O k-anonymity, no entanto, não impede ataques de inferência por homogeneidade ou ataques em que há conhecimento prévio de algum atributo sensível de determinado sujeito que conste no conjunto de dados, o que pode resultar na identificação desta pessoa (PATEL; AMIN, 2019). O objetivo do ataque por inferência é atribuir o valor de um atributo considerado como sensível para um conjunto de indivíduos, independentemente se estes indivíduos tenham sido reidentificados com sucesso ou não.

Por exemplo: dada uma base de dados de grupos anonimizados T* o atacante pode saber:

- Se dada pessoa tem uma trajetória anonimizada t* onde t* pertence a T*.
- Se todos os grupos g* pertencem a G*.
- Se dado grupo g* pertence a G*, uma pessoa p* pertence a g*.

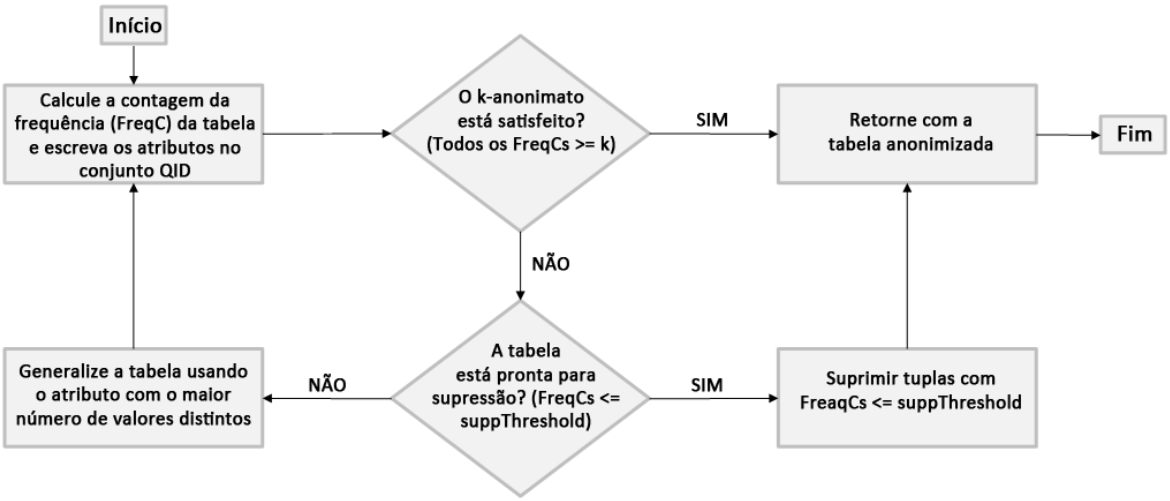
Com esses conhecimentos é possível realizar ataques sobre a base.

2.7 Algoritmos de k-anonynity

2.7.1 Datafly

É um algoritmo heurístico ganancioso que realiza generalização unidimensional de domínio completo (SWEENEY, 2002). Ele verifica a frequência do conjunto QID. Se k-anonymity não estiver satisfeito, ele generaliza o atributo com valores distintos até que a condição de k-anonymity seja satisfeita. Enquanto este algoritmo garante uma transformação k-anonymity, ele não fornece a generalização mínima (SAMARATI, 2001). A Figura 6 apresenta as etapas do Datafly.

Figura 6 – Adaptado de Processamento das iterações do algoritmo Datafly.



Fonte: (AYALA-RIVERA et al., 2014)

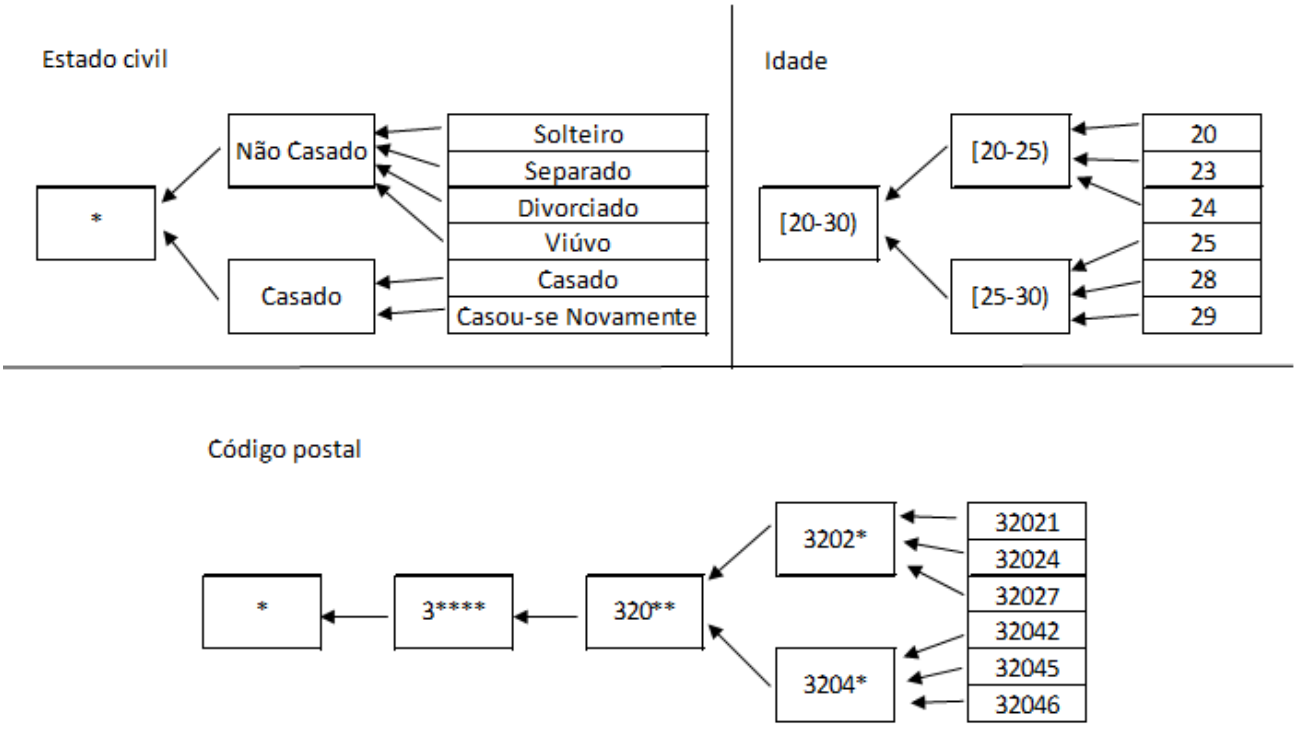
Abaixo são apresentados os dados para o processo de anonimização utilizando o algoritmo Datafly. As anonimizações intermediárias foram realizadas por iterações e a solução final apresentada. Os parâmetros de entrada para esta anonimização são: k = 2, suppThreshold = 0, QIDs = 3 (Estado Civil, Idade e Código Postal). A Tabela 6 apresenta os dados em seu formato original e as generalizações são guiadas pelos VGHs (Hierarquia de generalização de valor) apresentados na Figura 7.

Tabela 6 - Adaptado de Tabela de dados de antecedentes criminais.

	ID	Quase Identificadores (QIDS)			Atributo Sensível (SA)
Tuplas	Nome	Estado Civil	Idade	Código Postal	Crime
1	Joe	Separado	29	32042	Assassinato
2	Jill	Solteiro	20	32021	Roubo
3	Sue	Viúvo	24	32024	Tráfico
4	Abe	Separado	28	32046	Assalto
5	Bob	Viúvo	25	32045	Pirataria
6	Amy	Solteiro	23	32027	Abuso

Fonte: (AYALA-RIVERA et al., 2014)

Figura 7 – Adaptado de VGHs para os atributos: Estado Civil, Idade e Código Postal.



Fonte: (AYALA-RIVERA et al., 2014)

O algoritmo inicia verificando a contagem de frequência (FreqC) da Tabela inicial, para informar quantas ocorrências de cada item aparecem entre as iterações. Esse valor é incrementado à medida que novas ocorrências encontradas são iguais obedecendo os critérios de generalização.

Tabela 7 - Adaptado de Iteração 1 do algoritmo Datafly.

Quase Identificadores (QIDS)			
Estado Civil	Idade	Código Postal	FreqC
Separado	29	32042	1
Solteiro	20	32021	1
Viúvo	24	32024	1
Separado	28	32046	1
Viúvo	25	32045	1
Solteiro	23	32027	1

Fonte: (AYALA-RIVERA et al., 2014)

A Tabela 7 não satisfaz o parâmetro (k = 2 e QIDs = 3). Número de valores distintos por QID: Estado civil = 3 (Separado, Solteiro e Viúvo), Idade = 6 (20, 23, 24, 25, 28 e 29) e Código Postal = 6 (32042, 32021, 32024, 32046, 32045 e 32027). A Tabela precisa ser generalizada usando o atributo Idade.

Tabela 8 - Adaptado de Iteração 2 do algoritmo Datafly.

Quase Identificadores (QIDS)			
Estado Civil	Idade	Código Postal	FreqC
Separado	25-30	32042	1
Solteiro	20-25	32021	1
Viúvo	20-25	32024	1
Separado	25-30	32046	1
Viúvo	25-30	32045	1
Solteiro	20-25	32027	1

Fonte: (AYALA-RIVERA et al., 2014)

A Tabela 8 não satisfaz o parâmetro (k = 2 e QIDs = 3). Número de valores distintos por QID: Estado civil = 3 (Separado, Solteiro e Viúvo), Idade = 2 (25 - 30 e 20 - 25,) e Código Postal = 6 (32042, 32021, 32024, 32046, 32045 e 32027). A Tabela precisa ser generalizada usando o atributo Código Postal.

Tabela 9 - Adaptado de Iteração 3 do algoritmo Datafly.

Quase Identificadores (QIDS)			
Estado Civil	Idade	Código Postal	FreqC
Separado	25-30	3204*	1
Solteiro	20-25	3202*	1
Viúvo	20-25	3202*	1
Separado	25-30	3204*	1
Viúvo	25-30	3204*	1
Solteiro	20-25	3202*	1

Fonte: (AYALA-RIVERA et al., 2014)

A Tabela 9 não satisfaz o parâmetro (k = 2 e QIDs = 3). Número de valores distintos por QID: Estado Civil = 3 (Separado, Solteiro e Viúvo), Idade = 2 (25 - 30 e 20 - 25) e Código Postal = 2 (3204* e 3202*). A tabela possui frequências equivalentes e precisam ser agrupadas.

Tabela 10 - Adaptado de Iteração 3 do algoritmo Datafly.

Quase Identificadores (QIDS)			
Estado Civil	Idade	Código Postal	FreqC
Separado	25-30	3204*	2
Solteiro	20-25	3202*	2
Viúvo	20-25	3202*	1
Viúvo	25-30	3204*	1

Fonte: (AYALA-RIVERA et al., 2014)

A Tabela 10 precisa ser generalizada usando o atributo Estado Civil.

Tabela 11 - Adaptado de Iteração 4 do algoritmo Datafly.

Quase Identificadores (QIDS)			
Estado Civil	Idade	Código Postal	FreqC
Não-Casado	25-30	3204*	3
Não-Casado	20-25	3202*	3

Fonte: (AYALA-RIVERA et al., 2014)

A Tabela 11 satisfaz o parâmetro (k = 2 (mínimo de 2 registros iguais) e QIDs = 3). Número de valores distintos por QID: Estado Civil = 1 (Não Casado), Idade = 2 (25 - 30 e 20 - 25) e Código Postal = 2 (3204* e 3202*). A Tabela anonimizada é exibida como resultado final.

Tabela 12 - Adaptado de Tabela anonimizada.

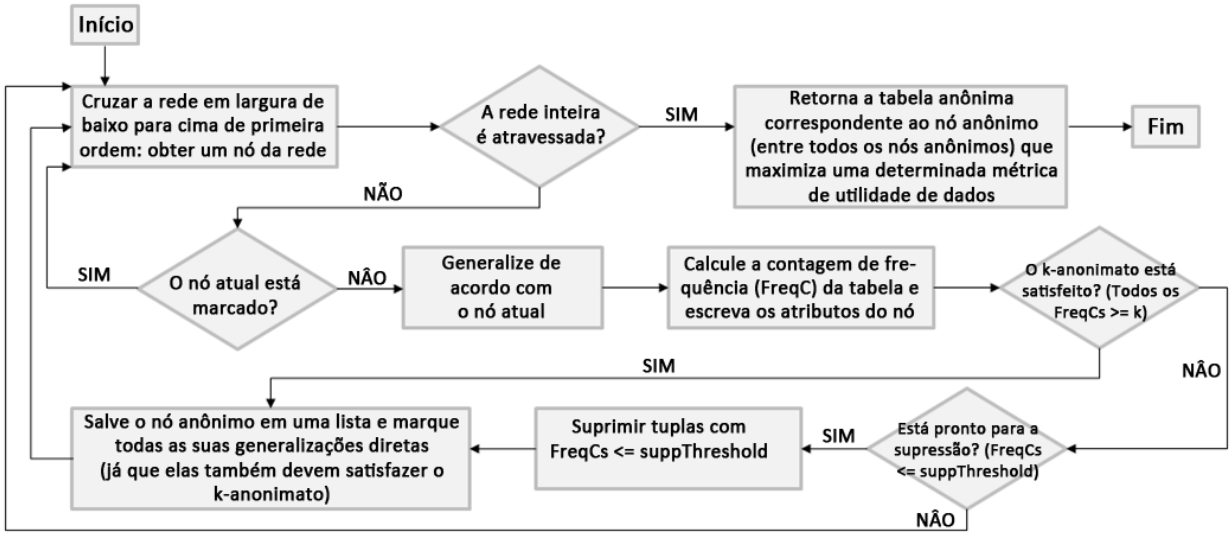
		Quase Identificadores (QIDS)			Atributo Sensível (SA)
Tuplas	EQ	Estado Civil	Idade	Código Postal	Crime
1	1	Não-Casado	25-30	3204*	Assassinato
4		Não-Casado	25-30	3204*	Assalto
5		Não-Casado	25-30	3204*	Pirataria
2	2	Não-Casado	20-25	3202*	Roubo
3		Não-Casado	20-25	3202*	Tráfico
6		Não-Casado	20-25	3202*	Abuso

Fonte: (AYALA-RIVERA et al., 2014)

2.7.2 Incognito

É um algoritmo unidimensional de generalização de domínio completo que constrói uma rede de generalização e a percorre usando uma busca abrangente de baixo para cima (LEFEVRE et al, 2005). A Figura 8 mostra as etapas do algoritmo.

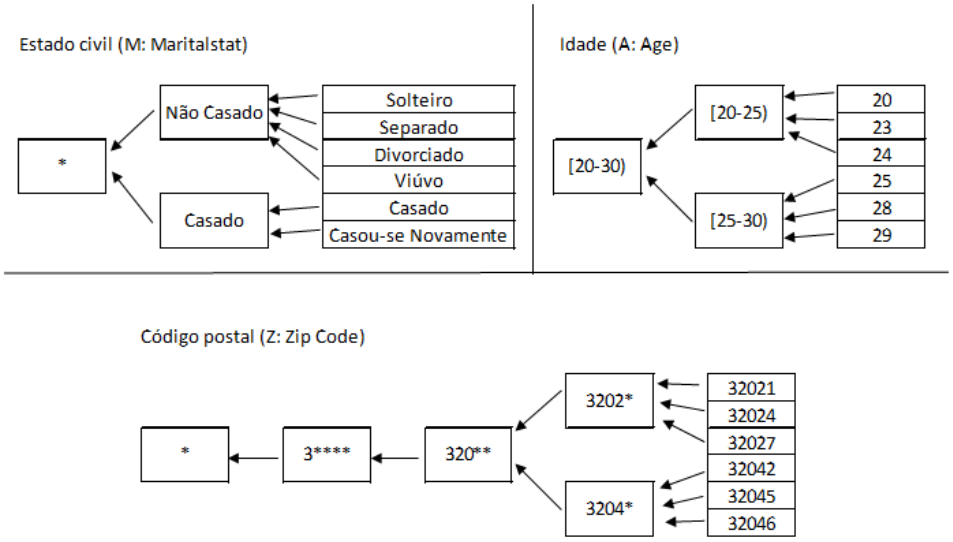
Figura 8 – Adaptado de Processamento das iterações do algoritmo Incognito.



Fonte: (AYALA-RIVERA et al., 2014)

A definição do número de generalizações que são consideradas válidas para um atributo é dado pela profundidade de seu VGH. Na Figura 9, seriam considerados: Estado Civil = 2, Idade = 2 e Código Postal = 4 considerando como base a Tabela 6 do item anterior. Como exemplo, a rede de generalização criada para esse conjunto QID.

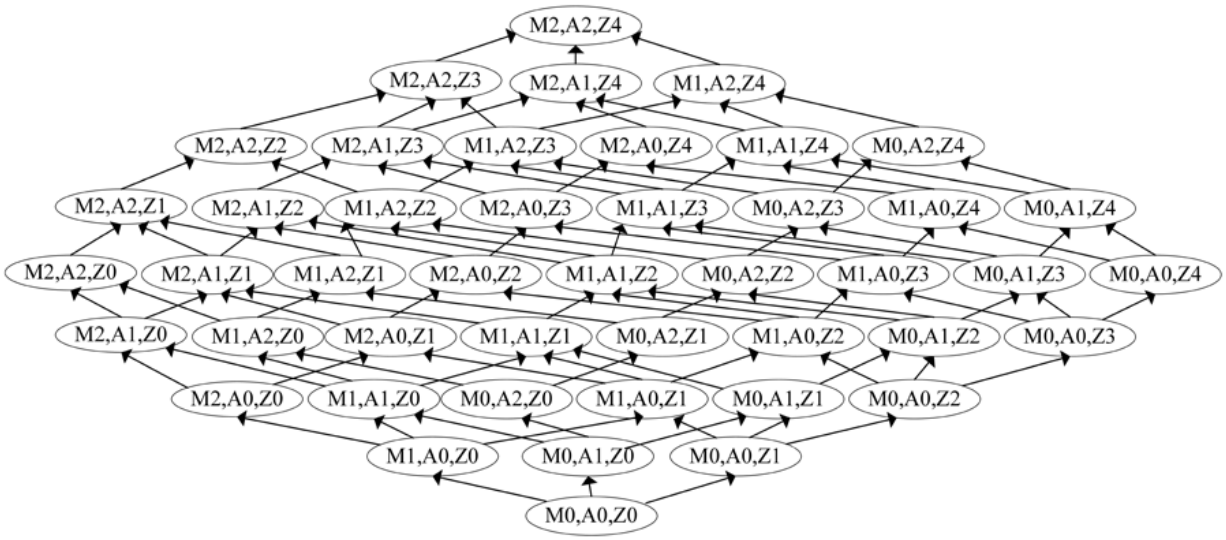
Figura 9 – Adaptado de Profundidade de VGHs dos atributos: Estado Civil: (M: MaritalStat), Idade: (A: Age) e Código Postal: (Z: ZipCode).



Fonte: (AYALA-RIVERA et al., 2014)

Os nós (estados de generalização disponíveis) são definidos pelos VGHs mostrados na Figura 9.

Figura 10 - Estados de generalização disponíveis.



Fonte: (AYALA-RIVERA et al., 2014)

Cada nó representa uma solução para o anonimato. Para transformar esses dados em dados anônimos, o algoritmo percorre a rede. Se um nó for encontrado e satisfazer o k-anonymity, todas as suas generalizações consideradas diretas, podem ser cortadas garantindo que satisfaça o k-anonymity. Diferentemente do Datafly, o Incognito cria uma solução anônima que contém a quantidade máxima de informações com base em uma métrica de informações selecionadas. Na implementação, o Incognito seleciona o resultado que produz o maior número de EQs. Sendo assim, a partir de uma coleção de soluções que atendem ao requisito de privacidade a solução final é escolhida satisfazendo k.

Para a Rede de Generalização da Figura 10, será apresentado dois caminhos que possuem solução para o k-anonimato. No primeiro nó (M1, A1, Z1) (Estado Civil (M) = 1, Idade (A) = 1, Código Postal (Z) = 1) que satisfazem o 2-anonimato, as três informações QIDs foram generalizadas um nível acima uma única vez. Essa solução é mostrada na Tabela 18. O segundo nó (M0, A2, Z2) (Estado Civil (M) = 0, Idade (A) = 2, Código Postal (Z) = 2) também satisfaz o 2-anonimato e as três informações QIDs foram generalizadas dois níveis acima. Essa solução é mostrada na Tabela 21.

Utilizando o modo de navegação anônima, a Tabela 13 é utilizada para ser anonimizada. Suas iterações são mostradas a seguir. Assim como o Algoritmo Datafly, o Incognito inicia verificando a contagem de frequência (FreqC) da Tabela inicial, para informar quantas ocorrências de cada item aparecem entre as iterações incrementado a contagem à medida que novas ocorrências encontradas são iguais obedecendo os critérios de generalização.

Tabela 13 - Adaptado de Tabela para ser anonimizada pelo Algoritmo Incognito.

Quase Identificadores (QIDS)			
Estado Civil	Idade	Código Postal	FreqC
Separado	29	32042	1
Solteiro	20	32021	1
Viúvo	24	32024	1
Separado	28	32046	1
Viúvo	25	32045	1
Solteiro	23	32027	1

Fonte: (AYALA-RIVERA et al., 2014)

Os nós da rede são verificados quanto ao k-anonymity. É apresentado também a solução final de anonimização, que é selecionada entre uma coleção de soluções que

satisfazem k, aquela que maximiza o número de EQs. Os parâmetros de entrada para esta anonimização são: k = 2, suppThreshold = 0, QIDs = 3 (Estado Civil, Idade e Código Postal).

O primeiro caminho a ser percorrido será: (M0, A0, Z0: nó inicial Tabela 13), (M0, A1, Z0), (M0, A1, Z1) e (M1, A1, Z1). A Tabela 13 não satisfaz o parâmetro (k = 2 e QIDs = 3). A Tabela precisa ser generalizada usando o atributo Idade.

Tabela 14 - Adaptado de Iteração do nó (M0, A1, Z0).

Verificação do nó (M0, A1, Z0)			
Quase Identificadores (QIDS)			
Estado Civil	Idade	Código Postal	FreqC
Separado	25-30	32042	1
Solteiro	20-25	32021	1
Viúvo	20-25	32024	1
Separado	25-30	32046	1
Viúvo	25-30	32045	1
Solteiro	20-25	32027	1

Fonte: (AYALA-RIVERA et al., 2014)

A Tabela 14 não satisfaz o parâmetro (k = 2 e QIDs = 3). A Tabela precisa ser generalizada usando o atributo Código Postal.

Tabela 15 - Adaptado de Iteração do nó (M0, A1, Z1).

Quase Identificadores (QIDS)			
Estado Civil	Idade	Código Postal	FreqC
Separado	25-30	3204*	1
Solteiro	20-25	3202*	1
Viúvo	20-25	3202*	1
Separado	25-30	3204*	1
Viúvo	25-30	3204*	1
Solteiro	20-25	3202*	1

Fonte: (AYALA-RIVERA et al., 2014)

A Tabela 15 possui frequências (registros) equivalentes e precisam ser agrupadas.

Tabela 16 - Adaptado de Iteração do nó (M1, A1, Z1).

Quase Identificadores (QIDS)			
Estado Civil	Idade	Código Postal	FreqC
Separado	25-30	3204*	2
Solteiro	20-25	3202*	2
Viúvo	20-25	3202*	1
Viúvo	25-30	3204*	1

Fonte: (AYALA-RIVERA et al., 2014)

A Tabela 16 não satisfaz o parâmetro (k = 2 e QIDs = 3). A Tabela precisa ser generalizada usando o atributo Estado Civil.

Tabela 17 - Adaptado resultado de Iteração do nó (M1, A1, Z1).

Quase Identificadores (QIDS)			
Estado Civil	Idade	Código Postal	FreqC
Não-Casado	25-30	3204*	3
Não-Casado	20-25	3202*	3

Fonte: (AYALA-RIVERA et al., 2014)

A Tabela 17 satisfaz (k = 2 e QIDs = 3). O resultado dessa Tabela anonimizada é apresentado na Tabela 18 e é marcada como uma possível solução final. Essa Tabela apresenta um EQ = 2.

Tabela 18 - Adaptado de Possível solução da Tabela 17 com o nó (M1, A1, Z1). Satisfaz o valor k.

	ID	Quase Identificadores (QIDS)			Atributo Sensível (SA)
Tuplas	EQ	Estado Civil	Idade	Código Postal	Crime
1	1	Não-Casado	25-30	3204*	Assassinato
4		Não-Casado	25-30	3204*	Assalto
5		Não-Casado	25-30	3204*	Pirataria
2	2	Não-Casado	20-25	3202*	Roubo
3		Não-Casado	20-25	3202*	Tráfico
6		Não-Casado	20-25	3202*	Abuso

Fonte: (AYALA-RIVERA et al., 2014)

Fazendo a varredura por outros nós da Figura 10, temos o segundo caminho: (M0, A0, Z0), (M0, A1, Z0), (M0, A1, Z1), (M0, A1, Z2) e (M0, A2, Z2). Serão apresentados somente os dois últimos nós (M0, A1, Z2) e (M0, A2, Z2) já que os nós (M0, A0, Z0), (M0, A1, Z0), (M0, A1, Z1) já foram percorridos e apresentados pelo caminho 1.

Tabela 19 - Adaptado de Iteração do nó (M0, A1, Z2).

Quase Identificadores (QIDS)			
Estado Civil	Idade	Código Postal	FreqC
Separado	25-30	320**	2
Solteiro	20-25	320**	2
Viúvo	20-25	320**	1
Viúvo	25-30	320**	1

Fonte: (AYALA-RIVERA et al., 2014)

Percorrendo pelo caminho 2, a Tabela 19 foi generalizada e agrupada usando o atributo Código Postal e ainda não satisfaz (k = 2 e QIDs = 3). A Tabela precisa ser generalizada usando o atributo Idade.

Tabela 20 - Adaptado de Iteração do nó (M0, A2, Z2).

Quase Identificadores (QIDS)			
Estado Civil	Idade	Código Postal	FreqC
Separado	20-30	320**	2
Solteiro	20-30	320**	2
Viúvo	20-30	320**	2

Fonte: (AYALA-RIVERA et al., 2014)

A Tabela 20 satisfaz (k = 2 e QIDs = 3). O resultado dessa Tabela anonimizada é apresentado na Tabela 21 e é marcada como uma possível solução final. Essa Tabela apresenta um EQ = 3.

Tabela 21 - Adaptado de Possível solução da Tabela 20 com o nó (M0, A2, Z2). Satisfaz o valor de k.

	ID	Quase Identificadores (QIDS)			Atributo Sensível (SA)
Tuplas	EQ	Estado Civil	Idade	Código Postal	Crime
1	1	Separado	20-30	320**	Assassinato
4		Separado	20-30	320**	Assalto
2	2	Solteiro	20-30	320**	Roubo
6		Solteiro	20-30	320**	Abuso
3	3	Viúvo	20-30	320**	Tráfico
5		Viúvo	20-30	320**	Pirataria

Fonte: (AYALA-RIVERA et al., 2014)

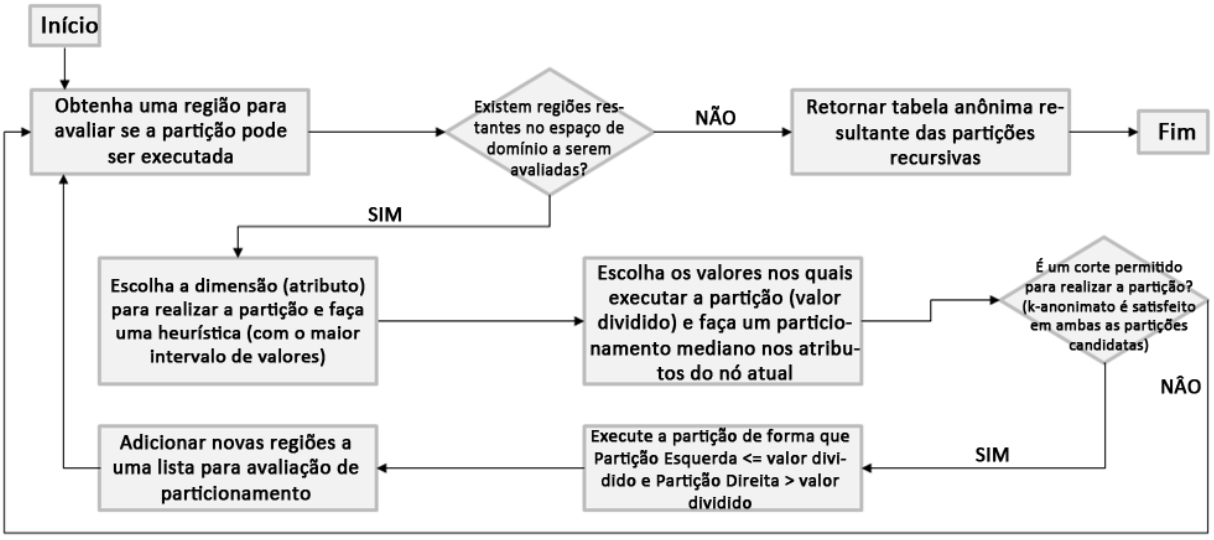
Uma vez que toda a rede foi percorrida, todos os nós resultantes devem ser analisados. Nas Tabelas 18 e 21 foram apresentados os resultados do processo de anonimização.

O resultado que satisfaz a condição do algoritmo Incognito como melhor solução é a Tabela 21 (caminho 2): nó (M0, A2, Z2). Ele produz o maior número de EQs dentre os resultados encontrados.

2.7.3 Mondrian

É um algoritmo multidimensional ganancioso que particiona o espaço do domínio recursivamente em várias regiões, cada uma das quais contém pelo menos k registros (LEFEVRE, et al, 2006). A Figura 11 apresenta as etapas do algoritmo de Mondrian.

Figura 11 – Adaptado de Processamento das iterações do algoritmo Mondrian.



Fonte: (AYALA-RIVERA et al., 2014)

O algoritmo inicia com um valor mais generalizado (menos especifico) dos atributos no conjunto QID. Há medida que as partições são executadas ele se especializa. Para escolher um determinado atributo onde realizará a partição, o algoritmo usa o atributo com intervalo normalizado de valores. Dentre as várias dimensões que possuem a mesma largura, será selecionada aquela que permite que um corte não cause uma violação do k-anonymity. Após a seleção, o algoritmo utiliza uma abordagem de particionamento mediana para definir o valor onde a partição será executada. Para encontrar a mediana de um atributo, uma abordagem de conjuntos de frequência é utilizada, ou seja, os dados são escaneados somando frequências para cada um dos valores únicos no atributo até que a posição mediana seja encontrada (LEFEVRE, et al, 2006). O valor localizado na mediana assume um valor dividido. No artigo original do Mondrian, o Mondrian já foi comparado com o Incognito, no entanto, esta comparação foi realizada apenas em termos da utilidade dos dados usando a métrica de discernibilidade (BAYARDO; AGRAWAL, 2005).

A Tabela 22 apresenta os atributos originais que serão utilizados para a criação da dimensão onde serão executados os cortes. Os atributos sensíveis serão apresentados na tabela final anonimizada.

Tabela 22 - Adaptado de Tabela original.

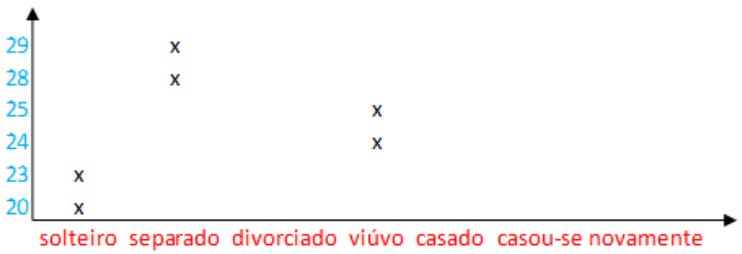
Quase Identificadores (QIDS)	
Estado Civil	Idade
Separado	29
Solteiro	20
Viúvo	24
Separado	28
Viúvo	25
Solteiro	23

Fonte: (AYALA-RIVERA et al., 2014)

Na Figura 12 é realizada a transformação da Tabela 22 utilizando o algoritmo Mondrian. No gráfico gerado pela transformação, é mostrado a dimensão geral da tabela representando os valores de QIDs. Cada ocorrência existente na tabela original é marcada um com X na representação gráfica. É apresentado nos gráficos seguintes algumas das partições recursivas realizadas e a solução final produzida. Para esse exemplo, foi selecionado apenas dois dos três atributos de QIDs. Os parâmetros de entrada são: k = 2 e QIDs = 2 (Idade e Estado Civil). Para o Estado Civil, foi utilizada a normalização de valores do parâmetro mapeando o atributo quanto ao seu VGH com a seguinte classificação: 1 - solteiro, 2 - separado, 3 - divorciado, 4 - viúvo, 5 - casado e 6 - casou-se novamente. Esse mapeamento é necessário, pois para realizar a mediana do tributo ele deve estar formatado na categoria numérico.

A heurística empregada neste exemplo para saber onde realizar as partições é a escolha da dimensão com o maior intervalo normalizado de valores. Da mesma forma, a estratégia de escolher o valor de divisão é executar o particionamento mediano.

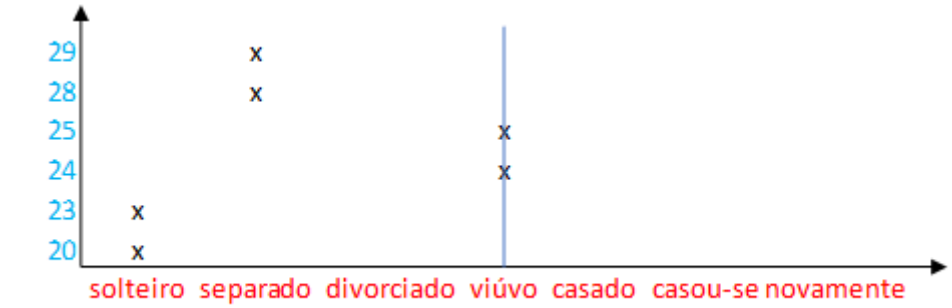
Figura 12 - Configuração inicial do algoritmo Mondrian.



Fonte: (O próprio autor)

Iteração 1: Região a ser avaliada: Região Completa - Idade (20 - 29) e Estado Civil (Solteiro - Casou-se novamente). Execute a partição fazendo a mediana no atributo Estado Civil. Mediana no intervalo (1 - 6) = 3,5. Não é permitido um corte com o valor encontrado.

Figura 13 - Iteração 1.



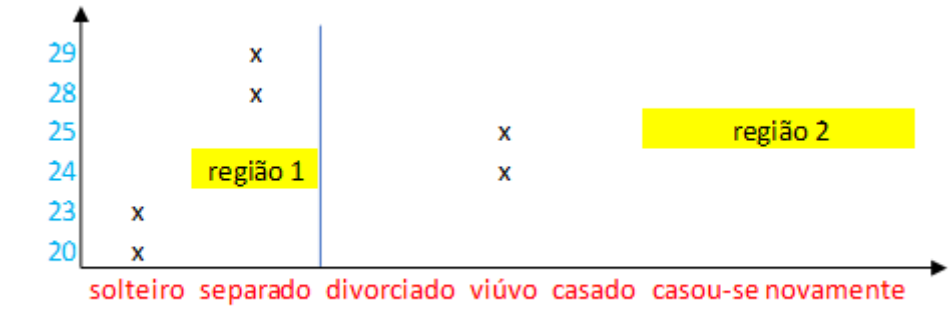
Fonte: (O próprio autor)

Calcule a mediana normalizando o último valor encontrado no intervalo anterior. Mediana no intervalo (1 - 3) = 2 (Separado). O valor de mediana encontrado para o Estado Civil é um corte permitido e não viola o valor de k.

Faça a partição no intervalo 2 (Separado). Partições resultantes:

- Partição esquerda (região 1): (20 - 29) - (Solteiro – Separado).
- Partição direita (região 2): (20 - 29) - (Separado - Casou-se Novamente)

Figura 14 - Continuação da Iteração 1.



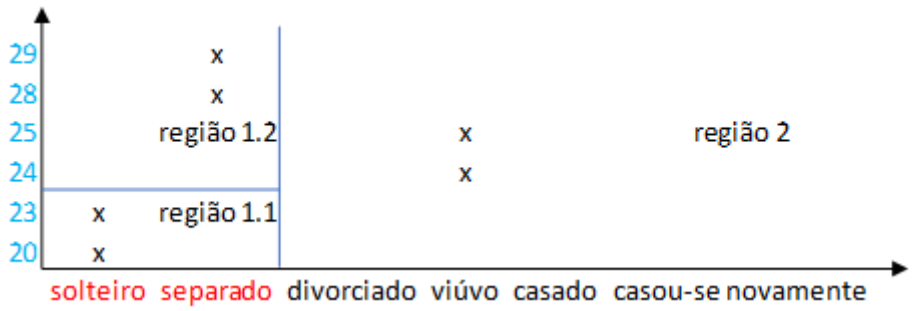
Fonte: (O próprio autor)

Iteração 2: Região a ser avaliada: Região 1 - Idade (20 - 29) e Estado Civil (Solteiro - Separado). Faça a mediana no atributo Idade. Mediana no intervalo (20 - 29) = 24,5. Não é permitido um corte com o valor encontrado. Calcule a mediana normalizando o último valor encontrado no intervalo anterior. Mediana no intervalo (20 - 24) = 23. O valor de mediana encontrado para Idade é um corte permitido e não viola o valor de k.

Partições resultantes:

- Partição esquerda (região 1.1): (20 - 23) - (solteiro - separado).
- Partição direita (região 1.2): (23 - 29) - (solteiro - separado).

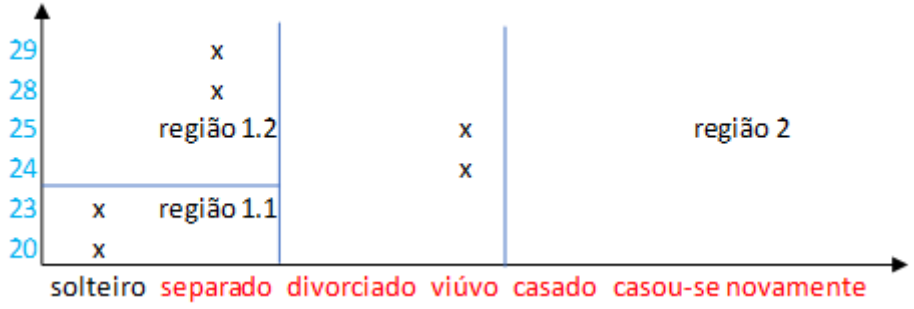
Figura 15 - Iteração 2.



Fonte: (O próprio autor)

Iteração 3: Região a ser avaliada: Região 2 – Idade (20 – 29) e Estado Civil (Separado – Casou-se novamente). Faça a mediana no atributo Estado Civil. Mediana no intervalo (2 - 6) = 4 (Viúvo). Não é permitido um corte com o valor encontrado, pois viola o valor de k em pelo menos uma das duas partições encontradas.

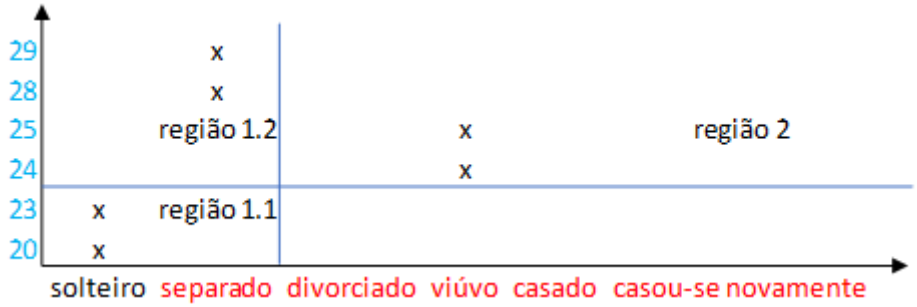
Figura 16 - Iteração 3.



Fonte: (O próprio autor)

Faça a mediana no atributo Idade. Mediana no intervalo (20 - 29) = 24,5. Não é permitido um corte com o valor encontrado. Calcule a mediana normalizando o último valor encontrado no intervalo anterior. Mediana no intervalo (20 - 24) = 23. O valor de mediana encontrado para Idade não é um corte permitido pois viola o valor de k, em pelo menos uma das duas partições encontradas.

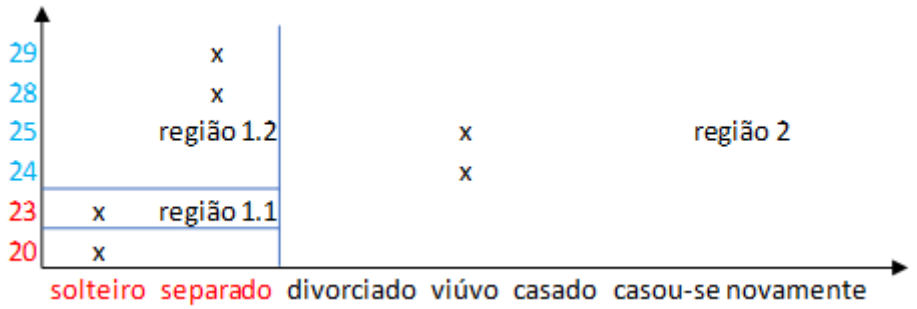
Figura 17 – Continuação da Iteração 3.



Fonte: (O próprio autor)

Não são permitidos cortes para a região 2. Prossiga com a análise das outras regiões. Iteração 4: Região a ser avaliada: Região 1.1 - Idade (20 - 23) e Estado Civil (Solteiro - Separado). Faça a mediana no atributo Idade. Mediana no intervalo (20 - 23) = 21,5. Não é permitido um corte com o valor encontrado. Calcule a mediana normalizando o último valor encontrado no intervalo anterior. Mediana no intervalo (20 - 21) = 20. O valor de mediana encontrado para Idade não é um corte permitido pois viola o valor de k em pelo menos duas partições encontradas.

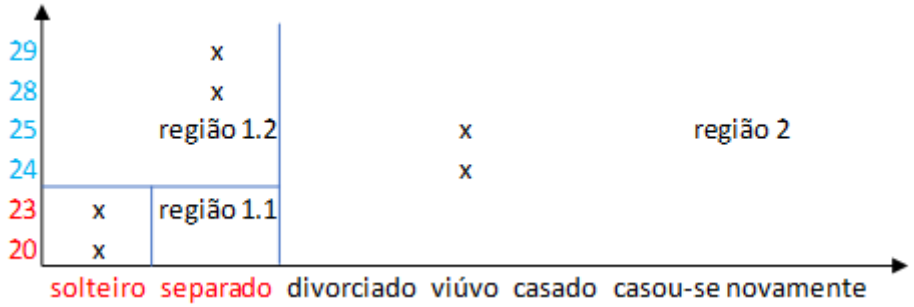
Figura 18 - Iteração 4.



Fonte: (O próprio autor)

Faça a mediana no atributo Estado Civil. Mediana no intervalo (1 - 2) = 1,5. Não é permitido um corte com o valor encontrado. Calcule a mediana normalizando o último valor encontrado no intervalo anterior. Mediana no intervalo (1) = 1 (Solteiro). Não é permitido um corte com o valor encontrado, pois viola o valor de k em uma das duas partições encontradas.

Figura 19 - Continuação da Iteração 4.

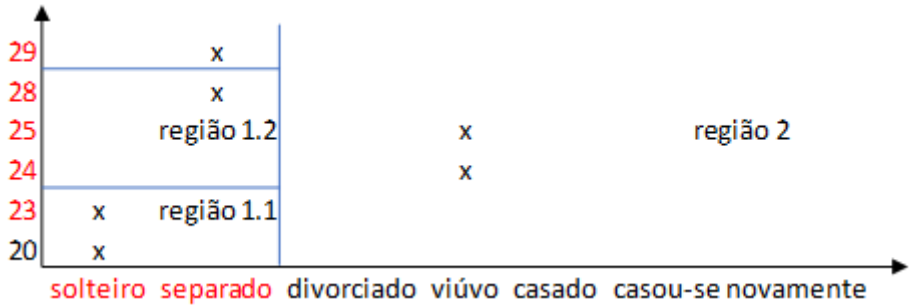


Fonte: (O próprio autor)

Não são permitidos cortes para a região 1.1. Prossiga com a análise das outras regiões.

Iteração 5: Região a ser avaliada: Região 1.2 - Idade (23 - 29) e Estado Civil (Solteiro - Separado). Faça a mediana no atributo Idade. Mediana no intervalo (23 - 29) = 25. Não é permitido um corte com o valor encontrado. Calcule a mediana normalizando o último valor encontrado no intervalo anterior. Mediana no intervalo (25 - 29) = 28. O valor de mediana encontrado para Idade não é um corte permitido pois viola o valor de k em pelo menos duas partições encontradas.

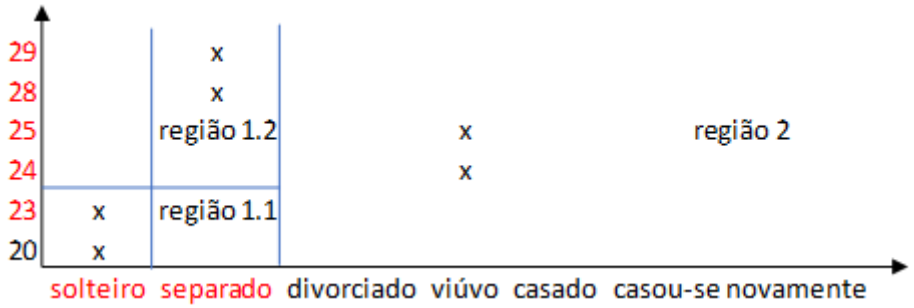
Figura 20 - Iteração 5.



Fonte: (O próprio autor)

Faça a mediana no atributo Estado Civil. Mediana no intervalo (1 - 2) = 1,5. Não é permitido um corte com o valor encontrado. Calcule a mediana normalizando o último valor encontrado no intervalo anterior. Mediana no intervalo (1) = 1 (Solteiro). Não é permitido um corte com o valor encontrado, pois viola o valor de k em uma das duas partições encontradas.

Figura 21 - Continuação da Iteração 5.



Fonte: (O próprio autor)

Não há mais regiões no espaço de domínio para serem avaliadas. A Tabela 23 apresenta os dados anonimizados.

Tabela 23 - Adaptado de Versão anônima da tabela 22.

	ID	Quase Identificadores (QIDS)			Atributo Sensível (SA)
Tuplas	EQ	Estado Civil	Idade	Código Postal	Crime
1	1	Solteiro-Separado	23-30	32042	Assassinato
4		Solteiro-Separado	23-30	32046	Assalto
2	2	Solteiro-Separado	20-23	32021	Roubo
6		Solteiro-Separado	20-23	32027	Abuso
3	3	Divorciado, Viúvo, Casado, Não-casado	20-30	32024	Tráfico
5		Divorciado, Viúvo, Casado, Não-casado	20-30	32045	Pirataria

Fonte: (AYALA-RIVERA et al., 2014)

2.8 Eficiência

Para realizar a tarefa de anonimização, o algoritmo deve ser analisado com base nos recursos exigidos para essa tarefa, pois esse processo pode exigir um alto consumo de recursos.

Dependendo da limitação desses recursos, algumas restrições podem afetar na escolha de um algoritmo. Se um algoritmo não for eficiente em uso de CPU, consumo de memória ou tempo de execução, mesmo se ele atingir um resultado satisfatório em utilidade nos dados, ele pode não ser considerado prático para uso.

Existem diversas maneiras de medir o tempo de execução do algoritmo. Pode-se levar em conta o upload dos dados para um banco de dados ou somente para a memória. Monitorar o tempo gasto nas etapas do processo de anonimização e também a saída dos dados. O processo de entrada e saída dos dados não mudam de um algoritmo para outro, sendo assim, será considerado somente o tempo de anonimização.

Para compararmos o desempenho do tempo de anonimização, será analisado o custo das seguintes características em cada um dos três algoritmos:

- DataFly: quantidade de operações de generalização executadas
- Incognito: estados de generalização na rede criada
- Mondrian: partições executadas nos dados

Outros recursos que também são bons indicadores de eficiência, são a memória e uso da CPU, apesar de dependerem da configuração do algoritmo. Implementações que fazem o carregamento dos dados na memória RAM, executam mais rápido o processo da anonimização, porém, para grandes volumes de dados, esses tipos de configuração não são escalonáveis, sendo necessário carregar os dados para um banco de dados. Nesta avaliação, como os algoritmos utilizam poucos recursos de CPU, não será considerado os valores de uso de CPU. Será considerado somente o uso de memória na execução das etapas do processo de anonimização.

Como as bases de dados podem ter tamanhos grandes, é necessário avaliar também a escalabilidade dos algoritmos. Dependendo da carga de execução planejada, é preciso estar ciente que a eficiência dos algoritmos reflete diretamente em grandes conjuntos de dados a serem anonimizados. Para esse comparativo, foi considerado também, a escalabilidade em relação ao tempo da anonimização e ao uso de memória.

2.8.1 Perda de Informação Generalizada (GenILoss)

De acordo com V. S. Iyengar (2002), esse parâmetro recebe a informação de perda ocorrida ao generalizar um atributo específico e ao quantificar a fração dos valores do domínio que foram generalizados. Para uma entrada ao atributo i, considere Li (limite inferior) e Ui (limite superior).

Essa entrada é generalizada para um intervalo ij definido pelos pontos iniciais de Lij inferior e Uij superior (NERGIZ; CLIFTON, 2007).

Para calcular a perda das informações, utiliza-se a fórmula abaixo:

$$GenILoss\ (T) = \frac{1}{|T| * n} * \sum_{i=1}^{n} \sum_{j=1}^{|T|} \frac{Uij - Lij}{Ui - Li}$$

- T: tabela com os dados originais
- n: número de atributos
- |T|: número de registros

O conceito por trás dessa métrica tem como base o fato de que valores com representatividade maior são menos expressivos do que valores com representatividade menor. Um exemplo prático desse conceito seria: não casado é menos expressivo do que divorciado ou solteiro. Valores baixos são mais expressivos:

- 0 nenhuma transformação (dados em seu formato original)
- 1 transformação total (nível máximo de generalização)

Apesar de ser utilizada somente em algoritmos que utilizam hierarquia de generalização, essa métrica será aplicada no algoritmo Mondrian que é não hierárquico. Qualquer faixa de valores generalizados pode ele, pode ser classificado dessa forma. Conforme afirma V. S. Iyengar (2002), ao aplicar a fórmula em atributos do tipo estado civil, o cálculo de penalidade mapeará cada dado para um valor numérico.

Por exemplo, considere a Tabela 24 com os dados originais onde são apresentados os atributos Estado Civil, Idade e Código Postal.

Tabela 24 - Tabela com dados originais.

Tabela Original		
Quase Identificadores (QIDS)		
Estado Civil	Idade	Código Postal
solteiro	20	32021
solteiro	23	32027
viúvo	24	32024
viúvo	25	32045
separado	28	32046
divorciado	29	32042

Fonte: (O próprio autor)

A Tabela 25 apresenta a anonimização da Tabela 24.

Tabela 25 - Dados da Tabela 24 anonimizados.

Tabela Anonimizada		
Quase Identificadores (QIDS)		
Estado Civil	Idade	Código Postal
não casado	20-25	3202*
não casado	20-25	3202*
não casado	20-25	3202*
não casado	25-30	3204*
não casado	25-30	3204*
não casado	25-30	3204*

Fonte: (O próprio autor)

Para exemplificar o método GenILoss, aplicaremos a fórmula na tabela 25 anonimizada.

Mapeando o atributo Estado Civil (atributo categórico convertido em numérico), na Tabela 26, temos o Limite Geral com base nos VGHs conhecidos: 1 - solteiro, 2 - separado, 3 - divorciado, 4 - viúvo, 5 - casado e 6 - casou-se novamente.

Tabela 26 - Limite Geral do atributo Estado Civil.

Estado Civil	VGHs		Limite Geral
não casado	solteiro	1	Inferior
	separado	2	
	divorciado	3	
	viúvo	4	
casado	casado	5	
	casou-se novamente	6	Superior
Tabela de classificação geral			

Fonte: (O próprio autor)

De acordo com esta classificação geral, o status de “não casado” apresenta o Limite Específico no intervalo de 1 até 4, que corresponde ao status de solteiro a viúvo.

Tabela 27 - Limite Específico do atributo Estado Civil.

Estado Civil	VGHs		Limite Específico
não casado	solteiro	1	Inferior
	separado	2	
	divorciado	3	
	viúvo	4	Superior
Tabela de classificação específica			

Fonte: (O próprio autor)

Ao aplicar a fórmula, temos as seguintes informações:

- Ui = 6 e Li = 1 (limite superior e inferior da Tabela 26 - Limite Geral).
- Uij = 4 e Lij = 1 (limite superior e inferior da Tabela 27 – Limite Específico).

Ao todo, temos 6 registros iguais com a informação “não casado”. De acordo com a fórmula, é necessário executar o somatório para todos esses registros.

$$\sum_{i=1}^{n}\sum_{j=1}^{|T|} = \frac{4-1}{6-1}+\frac{4-1}{6-1}+\frac{4-1}{6-1}+\frac{4-1}{6-1}+\frac{4-1}{6-1}+\frac{4-1}{6-1} = \frac{18}{5}$$

Para a Idade (atributo classificado como numérico), ao aplicar a fórmula em campos com valores “20-25” e “25-30”, temos as seguintes informações:

Mapeando o atributo Idade da tabela 26, temos o Limite Geral na Tabela 28.

Tabela 28 - Limite Geral do atributo Idade.

Idade	VGHs		Limite Geral
20-25	20	1	Inferior
	23	2	
	24	3	
25-30	25	4	
	28	5	
	29	6	Superior
Tabela de classificação geral			

Fonte: (O próprio autor)

Nas Tabelas 34 e 35, temos os Limites Específicos.

Tabela 29 - Limite Específico do atributo Idade 20-25.

Idade	VGHs		Limite Específico
20-25	20	1	Inferior
	23	2	
	24	3	Superior
Tabela de classificação específica			

Fonte: (O próprio autor)

Tabela 30 - Limite Específico do atributo Idade 25-30.

Idade	VGHs		Limite Específico
25-30	25	4	Inferior
	28	5	
	29	6	Superior
Tabela de classificação específica			

Fonte: (O próprio autor)

Para os 3 registros com “20-25”:

- Uij = 24 e Lij = 20 (limites superior e inferior da tabela 29 - Limite Específico)
- Ui = 29 e Li = 20 (limites superior e inferior da tabela 28 - Limite Geral).

Para os 3 registros com “25-30”:

- Uij = 29 e Lij = 25 (limites superior e inferior da tabela 30 – Limite Específico)
- Ui = 29 e Li = 20 (limites superior e inferior da tabela 28 - Limite Geral).

Executando o somatório da fórmula para todos os registros fica:

$$\sum_{i=1}^{n}\sum_{j=1}^{|T|} = \frac{24-20}{29-20}+\frac{24-20}{29-20}+\frac{24-20}{29-20}+\frac{29-25}{29-20}+\frac{29-25}{29-20}+\frac{29-25}{29-20}=\frac{24}{9}$$

Para o Código Postal (atributo categórico convertido em numérico), ao aplicar a fórmula em campos com valores “3202*” e “3204*”, temos as seguintes informações:

Mapeando o atributo Código Postal da Tabela 31, temos o Limite Geral:

Tabela 31 - Limite Geral do atributo Código Postal.

Código Postal	VGHs		Limite Geral
3202*	32021	1	Inferior
	32027	2	
	32024	3	
3204*	32045	4	
	32046	5	
	32042	6	Superior
Tabela de classificação geral			

Fonte: (O próprio autor)

Nas Tabelas 37 e 38, temos os Limites Específicos.

Tabela 32 - Limite Específico do atributo Código Postal 3202.*

Código Postal	VGHs		Limite Específico
3202*	32021	1	Inferior
	32027	2	
	32024	3	Superior
Tabela de classificação específica			

Fonte: (O próprio autor)

Tabela 33 - Limite Específico do atributo Código Postal 3204.*

Código Postal	VGHs		Limite Específico
3204*	32045	4	Inferior
	32046	5	
	32042	6	Superior
Tabela de classificação específica			

Fonte: (O próprio autor)

Para os 3 registros com “3202*”:

- Uij = 3 e Lij = 1 (limites superior e inferior da Tabela 32 - Limite Específico)
- Ui = 6 e Li = 1 (limites superior e inferior da Tabela 31 - Limite Geral).

Para os 3 registros com “3204*”:

- Uij = 6 e Lij = 4 (limites superior e inferior da tabela 33 - Limite Específico)
- Ui = 6 e Li = 1 (limites superior e inferior da tabela 31 - Limite Geral).

Executando o somatório da fórmula para todos os registros fica:

$$\sum_{i=1}^{n}\sum_{j=1}^{|T|} = \frac{3-1}{6-1}+\frac{3-1}{6-1}+\frac{3-1}{6-1}+\frac{6-4}{6-1}+\frac{6-4}{6-1}+\frac{6-4}{6-1}=\frac{12}{5}$$

Finalizando a fórmula com os campos restantes, a pontuação final do GenILoss para a tabela completa fica:

$$GenILoss\ (T) = \frac{1}{|T| * n} * \sum_{i=1}^{n} \sum_{j=1}^{|T|} = \frac{1}{6 * 3} * \frac{162 + 120 + 108}{45} = \frac{13}{27} = 0{,}48$$

Com base na referência abaixo:

- 0 (nenhuma transformação)
- 1 (transformação total)

o resultado encontrado apresenta um percentual aproximado de 50% de transformação dos dados.

2.8.2 Métrica de Dicernibilidade (Discernibility Metric - DM)

De acordo com R. J. Bayardo et, al (2005), essa métrica faz a distinção de um registro em relação a outros. Ao fazer esta distinção, ela aplica uma penalidade para o registro, igual ao tamanho do EQ (classes de equivalência) ao qual ele pertence. Caso um registro seja anulado, é atribuído uma penalidade igual ao tamanho da tabela de entrada.

Para calcular a pontuação geral de uma DM em uma tabela anônima, utiliza-se a formula abaixo:

$$DM\ (T) = \sum_{EQ} |EQ|^2$$

- T: tabela original
- |EQ|: tamanho de cada classe de equivalência. OBS: classes criadas após o processo de anonimização.

O conceito base dessa métrica, nos mostra que EQs maiores indicam perda de informação. Sendo assim, valores menores são considerados bons. Para exemplificar o método DM, aplicaremos a fórmula na Tabela 34.

Tabela 34 - Adaptado de Modelo de tabela anonimizada.

		Quase Identificadores (QIDS)			Atributo Sensível (SA)
Tuplas	EQ	Estado Civil	Idade	Código Postal	Crime
1	1	Não-Casado	25-30	3204*	Assassinato
4		Não-Casado	25-30	3204*	Assalto
5		Não-Casado	25-30	3204*	Pirataria
2	2	Não-Casado	20-25	3202*	Roubo
3		Não-Casado	20-25	3202*	Tráfico
6		Não-Casado	20-25	3202*	Abuso

Fonte: (AYALA-RIVERA et al., 2014)

Nessa tabela, os dois EQs listados possuem o tamanho 3. Calculando a pontuação dessa tabela, temos o seguinte resultado:

$$DM\ (T) = 3^2 + \ 3^2 = 18$$

2.8.3 Métrica do Tamanho de Classe de Equivalência (CAV G)

De acordo com K. LeFevre et, al (2005), essa métrica verifica como a criação de EQs, se aproximam do melhor caso, onde cada registro é generalizado em um EQ de k registros. O objetivo dessa métrica, é diminuir ao máximo a penalidade aplicada. O valor 1 indicaria a anonimização perfeita onde o tamanho dos EQs é dado pelo valor de k.

Para calcular a pontuação geral do CAVG em uma tabela anônima, utiliza-se a fórmula abaixo:

$$Cavg(T) = \frac{|T|}{|EQs| * k}$$

- T: tabela original
- |T|: número de registros
- |EQs|: total de classes de equivalência criadas
- K: requisito de privacidade

Calculando a pontuação CAV G para a tabela 12 com 2 EQs, temos o seguinte resultado:

$$Cavg(T) = \frac{6}{2 * 3} = 1$$

O resultado encontrado indica que a anonimização foi perfeita.

Capítulo 3 **METODOLOGIA**

A metodologia utilizada, consiste da preparação da base de dados e da configuração dos três algoritmos selecionados de anonimização (Datafly, Incógnito e Mondrian) utilizando as técnicas de generalização e supressão. Além disso, foram selecionadas algumas métricas, como por exemplo, Tempo de Anonimização, Consumo de Memória, Perda de Informação Generalizada (GenILoss), Métrica de Discernibilidade (DM) e Tamanho Médio de Classe de Equivalência (CAV G) para serem aplicadas nos resultados gerados pelos algoritmos para identificar qual deles pode ser considerado como um algoritmo ótimo para anonimização.

A base de dados escolhida foi do censo de adultos do UCI Machine Learning Repository (KOHAVI; BECKER 1996). Esta base apresenta bons dados para avaliar algoritmos de k-anonymity.

Os dados foram recebidos no formato .CSV e tiveram que ser convertidos para uma planilha com o objetivo de verificar a necessidade de formatação dos dados existentes. Foram encontrados no total, 30162 registros válidos.

A Tabela 35 apresenta as informações encontradas relacionadas a esses registros na base de dados. Nela, são listados os atributos desses registros, seus números de valores distintos e a altura VGH (hierarquia de generalização de valor) atribuída a cada atributo. Na coluna final, é especificado o número de níveis de generalizações disponíveis que podem ser utilizados durante a execução dos algoritmos para cada atributo.

Tabela 35 - Conjunto de dados adultos.

Id	Atributo	Valores	Generalizações (Tamanho VGH)
1	Idade	74	4 no intervalo (5, 10, 20) anos
2	Gênero	2	1
3	Raça	5	1
4	Estado Civil	7	2
5	País Nativo	41	2
6	Classe de Trabalho	8	2
7	Ocupação	14	2
8	Escolaridade	16	3
9	Salário	2	1

Fonte: (KOHAVI; BECKER (1996))

De acordo com V. S. Iyengar (2002), dados como esses, precisam ser analisados e preparados removendo todas as entradas com valores ausentes ou nulos, bem como possíveis erros de conversão de dados deixando um padrão para utilização posterior. Ao realizar essa análise, percebeu-se que alguns registros apresentaram falhas no processo de conversão da tabela original. Exemplo: alguns registros como “Never@married” tiveram que ser corrigidos para o formato “Never-married”. Outras falhas encontradas, foram caracteres especiais no final do registro, como por exemplo, “Private?”. Para esses casos, a simples exclusão desses caracteres especiais “?” foi o suficiente para ter os registros corrigidos. Isso não dificultou em preservar a padronização das informações. No final da tabela apareceram campos vazios que foram deletados, mantendo-se somente os campos contendo os registros.

Após toda a análise, verificação da padronização dos dados e as devidas correções, a base de dados foi novamente convertida para seu formato original para ser utilizada nos algoritmos.

O motivo da seleção dos algoritmos se deve pelos seguintes motivos:

- Esses algoritmos são bastante citados na literatura
- Existe implementação pública desses algoritmos disponível
- As diferentes estratégias de anonimização desses algoritmos permitem uma avaliação mais ampla
- Podem ser analisados dentro de uma estrutura possibilitando uma comparação coerente

- Pode ser avaliado o grau de eficiência de cada algoritmo obedecendo os seguintes parâmetros: Tempo de execução total de cada algoritmo, consumo de memória, tempo do processo de anonimização das diferentes etapas de cada algoritmo, perda de dados durante o processo de anonimização e a relação do tempo de anonimização e o custo das características funcionais de cada algoritmo.

A partir de uma definição comum, é fundamental analisar os algoritmos tendo como base os parâmetros usados em suas avaliações anteriores. Realizar comparações, exige um alto grau de critérios que possam ser aplicados para analisar diversos aspectos (por exemplo, utilidade dos dados e eficiência).

Deve-se levar em conta a quantidade de recursos para realizar os procedimentos de anonimização, pois todo o procedimento deve focar no consumo desses recursos. São consideradas as seguintes etapas: a entrada dos dados, o procedimento de anonimização e o resultado final fornecidos pelas métricas.

Para a bateria de testes, no caso de desempenho dos algoritmos, foi verificado a relação entre custos funcionais e tempo de anonimato com base nas seguintes características de cada algoritmo:

- DataFly: Quantidade de generalizações realizadas
- Incognito: Estados de generalização ou número de nós
- Mondrian: Quantidade de partições nos dados

Para a realização da comparação dos algoritmos, todos eles foram executados em sua forma original, alterando, somente, alguns parâmetros de entrada. Cada algoritmo possui configurações e métricas diferentes. O desempenho poderá variar devido as diversas combinações de parâmetros de entrada e conjuntos de dados. Dependendo do cenário de comparação, cada um pode funcionar bem em certos cenários e funcionar mal em outros.

Para um resultado justo, é importante considerar uma configuração padrão entre eles que reflita diretamente nos parâmetros utilizados nas avaliações. Uma comparação exige o uso alguns critérios que podem servir para medir diferentes aspectos como a utilidade de dados e eficiência.

Como os cenários são desconhecidos, além do tempo de anonimização e consumo de memória, em termos de utilidades de dados, serão considerados as seguintes métricas:

- Perda de informação generalizada (GenILoss). Esse parâmetro recebe a informação de perda ocorrida ao generalizar um atributo específico que pode ser aplicada em uso geral de diversos cenários.
- Métrica de Discernibilidade (DM). Essa métrica faz a distinção de um registro em relação a outros.
- Tamanho Médio de Classe de Equivalência (CAV G). Essa métrica verifica como a criação das classes de equivalência (EQs), se aproximam do melhor caso.

Capítulo 4 EXPERIMENTOS

Para a execução dos experimentos, foi utilizada uma máquina com o sistema operacional Ubuntu 20.04 LTS de 64 bits, processador Intel Core I7-10700K com clock de 3,80 GHz e 16 GB de memória RAM. Foram apresentados por Ayala-Rivera et al. (2014), os experimentos que serviram de base para a execução do nosso comparativo. Nos experimentos originais, foram utilizadas algumas aplicações desenvolvidas pelos próprios autores juntamente com o suporte de um banco de dados MySQL 5.5.34 como base para armazenar estados intermediários de anonimato. Para nossos experimentos, optamos por recriar esses experimentos enviando todas as informações direto para a memória RAM uma vez que a quantidade de registros contidos na base de dados, podem facilmente serem processadas. A ferramenta utilizada para executar o processo de anonimização na base de dados com os três algoritmos foi a UT Dallas Anonymization Toolbox (KANTARCIOGLU, 2019). Os parâmetros utilizados estão descritos na Tabela 36.

Tabela 36 - Configuração dos experimentos.

Experimentos	Parâmetros	Registros
1 - Variando QIDs	K = 2 QIDs = [1, 2, 3, 4, 5]	30162
2 - Variando K	k = [2, 5, 10, 25, 50, 100, 250, 500, 1000] QIDs= 3	
3 - Variando Registros	K = 50 e QIDs = 3	[5.000, 10.000, 15.000, 20.000, 25.000]

Fonte: (O próprio autor)

A variação dos parâmetros feita nesses experimentos foram:

- |QIDs|: número de atributos utilizados. Para o experimento 1, foram considerados: Gênero, Idade, Raça, Estado Civil e País Nativo. Para o experimento 2, Foram considerados: Estado Civil, Idade e País Nativo.

- Valor de K: nível de privacidade que deve ser realizado no processo de anonimização.
- Variando o tamanho do arquivo: experimento extra alternando a quantidade de registros presentes no arquivo.

Foi utilizada essa variação nos valores de K e dos QIDs nos experimentos para verificar se os algoritmos apresentam alguma mudança de comportamento quando for aplicado uma alteração nos parâmetros de anonimização, bem como foi adicionado um experimento extra para verificar se o tamanho do arquivo também influência nos resultados.

Essas variações contribuem para um melhor entendimento dos resultados e possibilita entender melhor como essas alterações pode influenciar na execução dos algoritmos testados.

Durante o procedimento de anonimização, todos os registros formatos foram considerados. Com base nos parâmetros informados, serão apresentados os resultados obtidos, sendo discutido as métricas relevantes para cada um dos experimentos. Serão considerados como resultados ótimos, os valores mais baixos encontrados em todas as métricas.

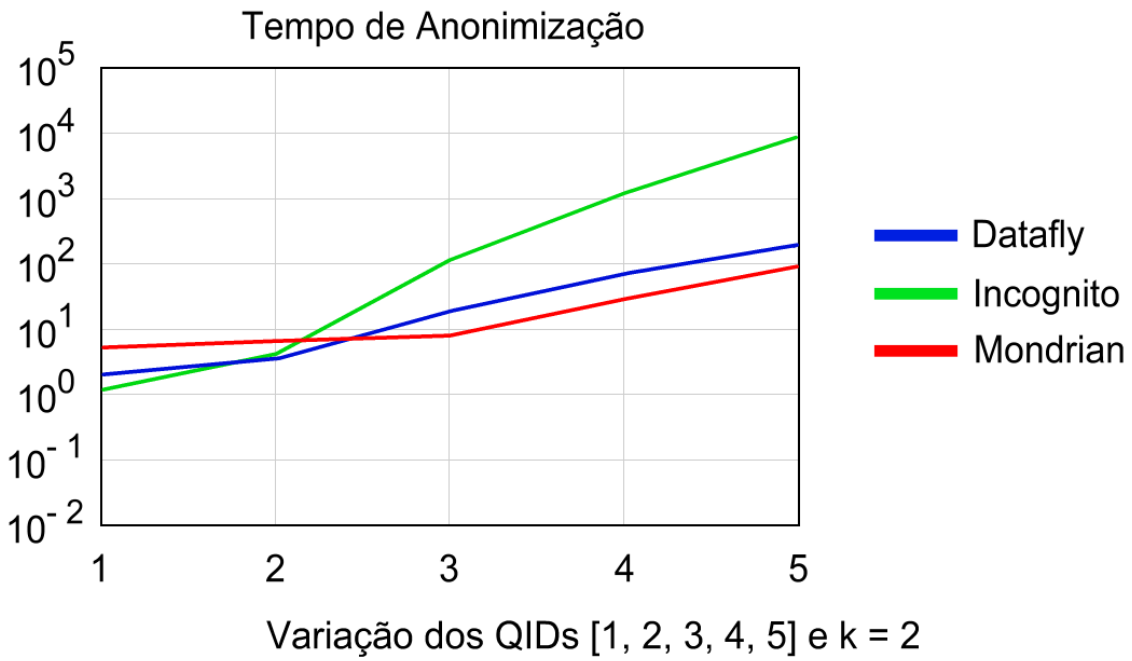
4.1 Experimento 1: Variação de QIDs

Para este experimento, é utilizado o termo QIDs (quase identificador), para indicar o número de atributos que compõem o conjunto de identificadores. O tamanho de um QID é definido com base nos primeiros atributos rotulados na base de dados utilizada. O desempenho dos algoritmos foi medido à medida que o número de QIDs vai aumentando dentro do intervalo {1, 2, 3, 4, 5}.

4.1.1 Tempo de Anonimização

O gráfico da Figura 22, apresenta o tempo gasto (em segundos) pelos algoritmos para realizar a anonimização conforme cresce o número de QIDs (quase identificadores).

Figura 22 - Tempo de anonimização experimento 1.



Fonte: (O próprio autor)

O algoritmo Mondrian, mesmo se a base de dados já satisfazer o k-anonimato, realiza a anonimização dos dados por não utilizar como critério de parada o k-anonimato. Ele executa todas as possibilidades de partições de maneira que não seja mais permitido realizar cortes.

Já o Incognito e o Datafly, apresentam um tempo de anonimização quase 0 quando seu QIDs pertence a {1,2}. Isso acontece por que eles já satisfazem a configuração 2-anonimato para esta base de dados não sendo necessário realizar nenhuma generalização. Com exceção dos experimentos onde nenhuma generalização foi executada, o incógnito apresenta o pior desempenho de todos os algoritmos testados.

Quando os QIDs são maiores que 2, o Incognito mostra-se muito mais lento que o Datafly seguido do Mondrian. Esse desempenho de anonimização acontece porque, o algoritmo, procura soluções para executar essa anonimização conforme mais atributos vão sendo inseridos no conjunto QID, pois ele precisa percorrer a rede para generalizar e analisar o estado de k-anonimato para cada condição individual.

O parâmetro que determina qual o número total de nós que devem ser avaliados depende da profundidade dos VGH (hierarquia de generalização de valor) definidos para cada atributo individual. Isso faz com que o tempo de anonimização do Incognito tenha um grande aumento.

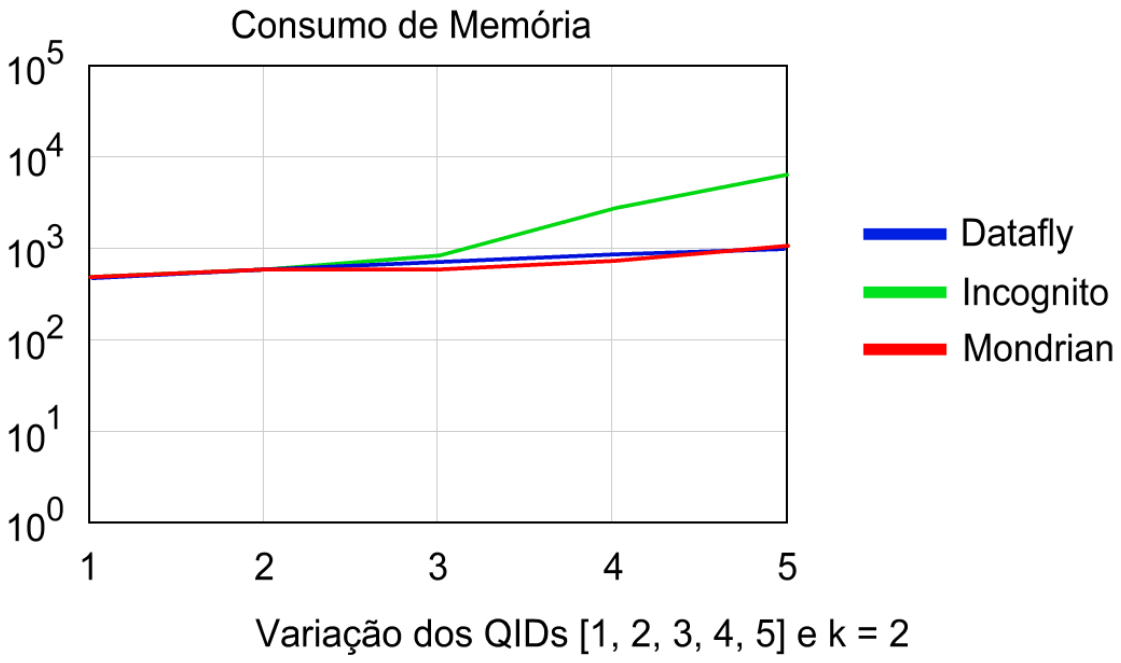
Percebe-se essa variação quando o número de QIDs é maior que 2. Somente quando o número de QID é menor, o incógnito apresenta bom desempenho. Além do número de atributos do conjunto QID, outro fator que influencia diretamente no aumento do tempo de anonimato do Incognito é a quantidade de estados possíveis de generalização que cada atributo pode ter.

Quando o QIDs é igual a 5, o Datafly e o Mondrian apresentam um desempenho bem próximo para concluir o processo de anonimização. Esse desempenho é considerado aceitável.

4.1.2 Consumo de Memória

O gráfico da Figura 23, mostra o consumo de memória dos 3 algoritmos conforme o número de QIDs aumenta.

Figura 23 - Consumo de memória experimento 1.



Fonte: (O próprio autor)

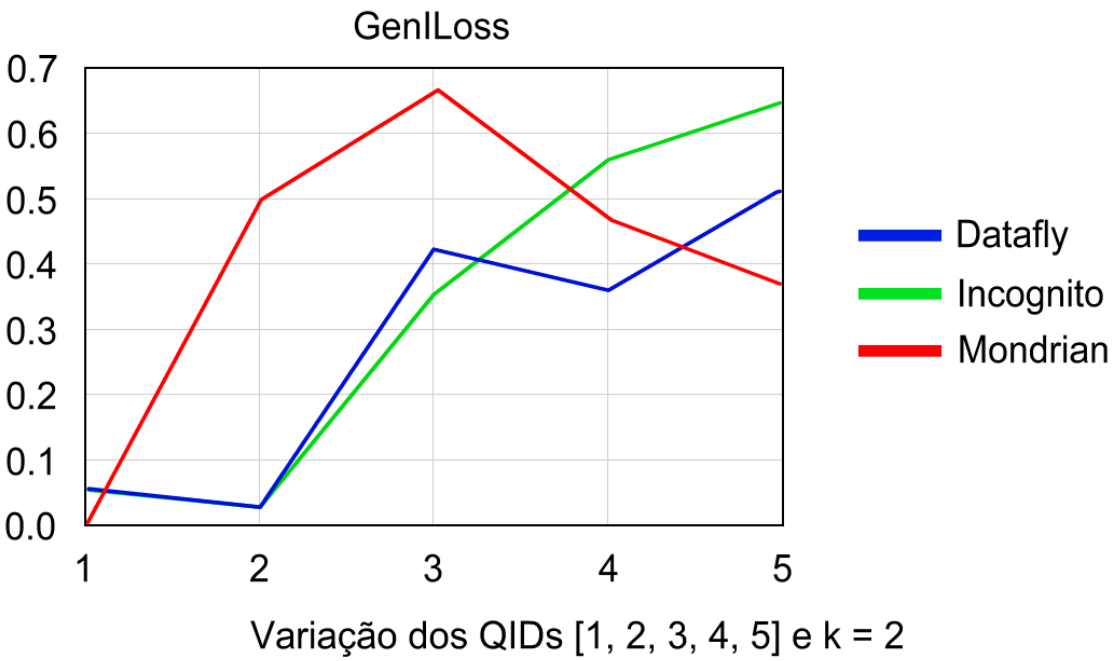
O incógnito, conforme observado no gráfico, apresenta um consumo maior de memória devido à grande quantidade de nós utilizados no processo de generalização. Esse consumo se expande à medida que o número de QIDs aumenta. Já o Mondrian e o Datafly, mostram

uma pequena variação desse consumo e apresentam uma condição estável à medida que o número de QIDs aumenta.

4.1.3 Perda de Informação Generalizada (GenILoss)

O gráfico da Figura 11.2, mostra os resultados da métrica GenILoss para os utilitários de dados conforme aumenta o número de QIDs.

Figura 24 - GenILoss experimento 1.



Fonte: (O próprio autor)

Os algoritmos Mondrian e Datafly apresentam um comportamento que não tem uma definição padrão. Eles retornam resultados distintos. Percebe-se que à medida que o número de QIDs aumenta, a perda de informação tende a ser alta.

O Mondrian apresenta um resultado muito ruim quando o QIDs está entre {2,3}. Isso acontece pelo fato de o Mondrian particionar os dados utilizando um único atributo. Com isso, os valores que são considerados menos específicos, são retidos no conjunto QID gerando uma grande penalidade para os atributos. Por exemplo, considere um QIDs = 3 que utiliza os seguintes atributos: idade, gênero e raça. Desses 3 atributos, a idade é o único atributo que permite realizar um corte, pois, de acordo com as regras de iteração do Mondrian, os outros atributos não satisfazem a partição média e devem ser descartados. Na primeira iteração do

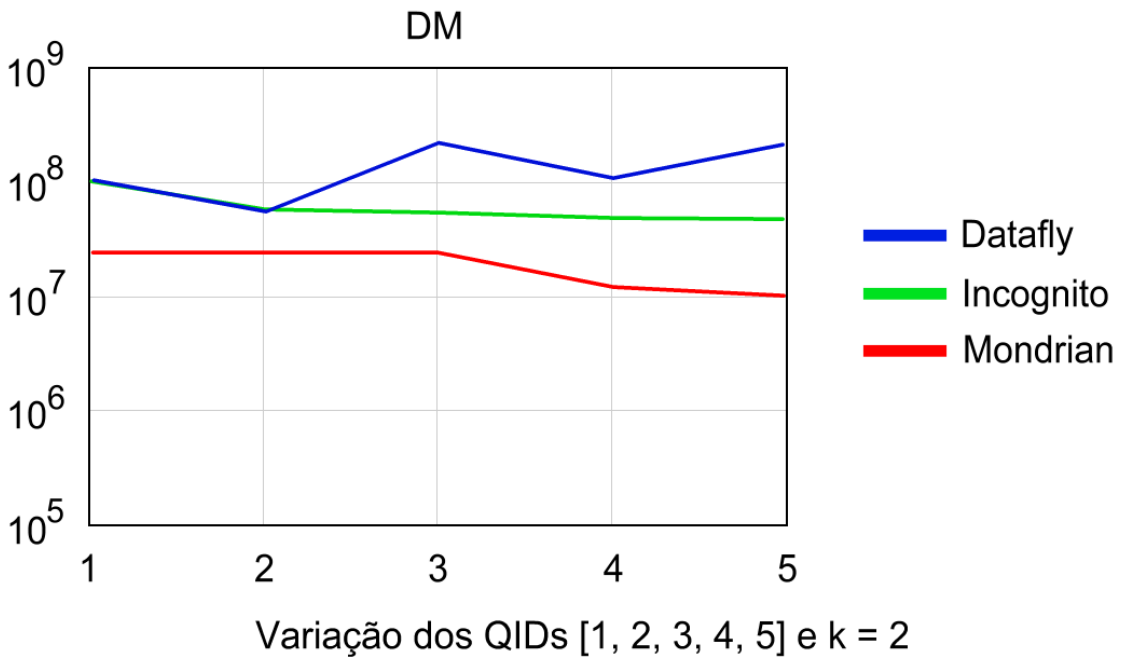
Mondrian, quando os dados são lidos para determinar o valor mediano para o tributo gênero, esse atributo apresenta 2 valores: feminino com 9.782 registros e masculino com 20.380 registros. Se considerarmos todos os registros que é permitido quando a frequência do gênero masculino é somada à contagem, atingimos a mediana em 15.081. Se tentar realizar um corte nesse valor, não será permitido. Para que seja possível fazer o corte, deve ter um tamanho mínimo de registros para cada nova partição que satisfaça o valor de k. Sendo assim, não satisfaz a condição. Seus valores retém o estado menos específico não podendo ser particionados utilizando esse atributo. Esse comportamento acontece em outras iterações independentemente do nível em que se encontra, pois, a distribuição dos dados é a mesma. O objetivo da técnica de partição mediana é conseguir uma posição uniforme. Quando os dados são distorcidos, essa técnica gera uma grande perda de informação. Isso comprova que o Mondrian trabalha melhor em condições uniformes onde a maioria das QIDs podem ser usadas para particionar dados. Do contrário, o Mondrian apresentaria um resultado muito alto no GenILoss. Atributos não usados em particionamento tem seus valores menos específicos retidos.

Os algoritmos Incognito e Datafly, tem tendências a ter o seu valor de GenILoss alto. Isso se deve pelo fato da maioria dos atributos serem generalizados mais vezes para que o k-anonimato seja atingido, degradando a utilidade dos dados.

4.1.4 Métrica de Discernibilidade (DM)

O gráfico da Figura 25, mostra os resultados da métrica DM para os utilitários de dados conforme aumenta o número de QIDs.

Figura 25 - DM experimento 1.



Fonte: (O próprio autor)

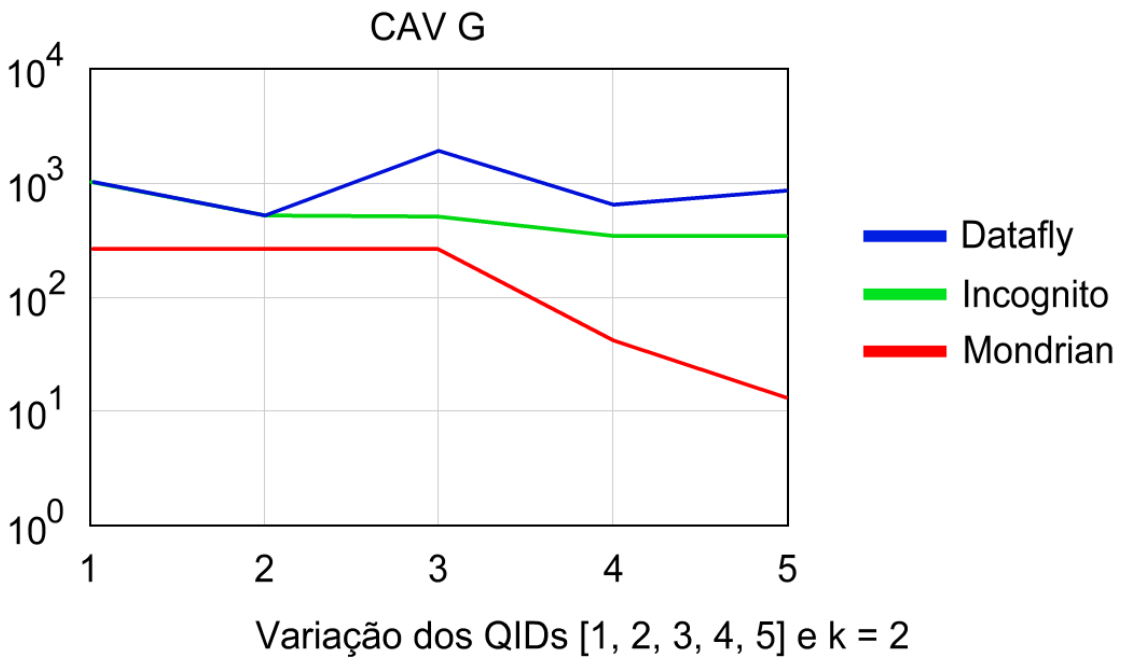
O Mondrian apresenta melhor desempenho em relação ao Incognito e Datafly. Isso se deve ao fato de que o algoritmo cria EQs de tamanhos pequenos. Outro ponto importante é que o DM não retém as transformações ocorridas nos QIDs. Quando o QIDs do Mondrian está em {1,3}, obtém-se o mesmo valor de DM, conforme mostrado no gráfico. Isso acontece, por que o Mondrian cria partições com um único atributo, conforme já mencionado. Sendo assim, são criados os mesmos EQs nos 2 experimentos. Sempre que o número de QIDs aumenta (adição de novos atributos), o Mondrian cria novas partições e com isso melhora o valor recebido na métrica DM. Um possível problema que pode piorar diretamente o resultado do DM no Mondrian, é quando vários atributos possuem o mesmo intervalo de valores e são usados critérios para selecionar o atributo com a maior quantidade de valores onde a partição será realizada. No experimento, o algoritmo selecionou o primeiro atributo. Essa condição reduz o número de partições em alguns cenários, pois afeta diretamente nas partições criadas.

O Incognito e o Datafly, se mostram iguais quanto ao DM, quando o valor do QIDs está em {1,2}. Isso se deve ao fato de que ambos atingem a mesma solução de anonimato. Como o Incognito encontra a generalização que produz um número maior de EQs, ele se mostra melhor que o Datafly, considerando um número maior de QIDs. Não se pode dizer o mesmo em relação ao Mondrian, pois o Mondrian tem como grande benefício uma abordagem multidimensional.

4.1.5 Tamanho Médio de Classe de Equivalência (CAV G)

O gráfico da Figura 26, mostra os resultados da métrica CAV G para os utilitários de dados conforme aumenta o número de QIDs.

Figura 26 - CAV G experimento 1.



Fonte: (O próprio autor)

Essa métrica apresenta um resultado gráfico bem parecido ao resultado obtido no DM, porém com outras proporções.

De acordo com o gráfico, o Mondrian se mostrou melhor em todos os QIDs nessa métrica. Isso se deve aos números de EQs criados no procedimento de anonimização. De modo geral, quando pelo menos dois algoritmos possuem a mesma pontuação de CAV G, significa que eles produziram o mesmo número de EQs, porém, não é garantido que esses EQs, possuem exatamente o mesmo tamanho. Isso comprova que a métrica CAV G, não retém a distribuição de registros entre os EQs.

4.1.6 Resumo do Experimento 1

Em termos de tempo de anonimização e consumo de memória, os algoritmos Mondrian e Datafly apresentaram melhor desempenho do que o Incognito. Pelo tamanho do conjunto de QIDs, o tempo gasto é considerável razoável para ambos. Apesar do Mondrian e Datafly exibirem um crescimento exponencial, a magnitude do Incognito é maior se comparado aos dois. Além dos inúmeros estados de anonimização influenciarem no tempo, o número de QIDs também alteram o resultado dessa métrica. Baseando-se nos utilitários de dados DM e CAV G, quanto ao tamanho dos EQs, o Mondrian supera os outros algoritmos. Quanto ao GenILoss, o Mondrian é penalizado por outliers, pois, particionamento mediano não pode ser executado para alguns atributos, fazendo com que o resultado do GenILoss atinja valores altos.

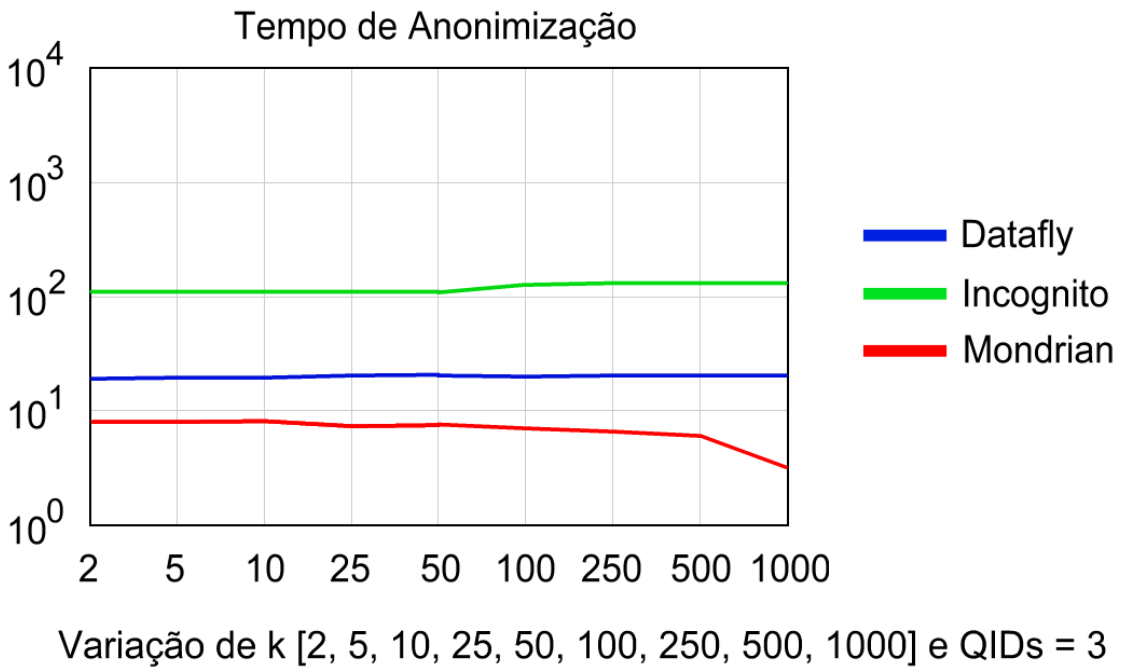
4.2 Experimento 2: Variação de K

Para este experimento, será usado uma variação de k com os seguintes valores: {2, 5, 10, 25, 50, 100, 250, 500, 1000,} e um | QIDs | = 3.

4.2.1 Tempo de Anonimização

O gráfico da Figura 27, apresenta o tempo gasto (em segundos) pelos algoritmos para realizar a anonimização conforme à medida que o valor de k aumenta.

Figura 27 - Tempo de anonimização experimento 2.



Fonte: (O próprio autor)

Conforme o valor de k vai aumentando, aumenta-se também a quantidade de generalizações necessárias para satisfazer o k-anonimato. Com esses fatores, o tempo necessário para realizar o anonimato aumenta conforme a variação do valor k. Isso se deve, pelo fato de satisfazer níveis mais altos de privacidade, é mais difícil.

O Incognito apresenta uma sensibilidade reduzida à medida que o valor de k aumenta.

Para o Datafly, a diferença de tempo é quase inexistente sendo considerado estável em praticamente todos os cenários.

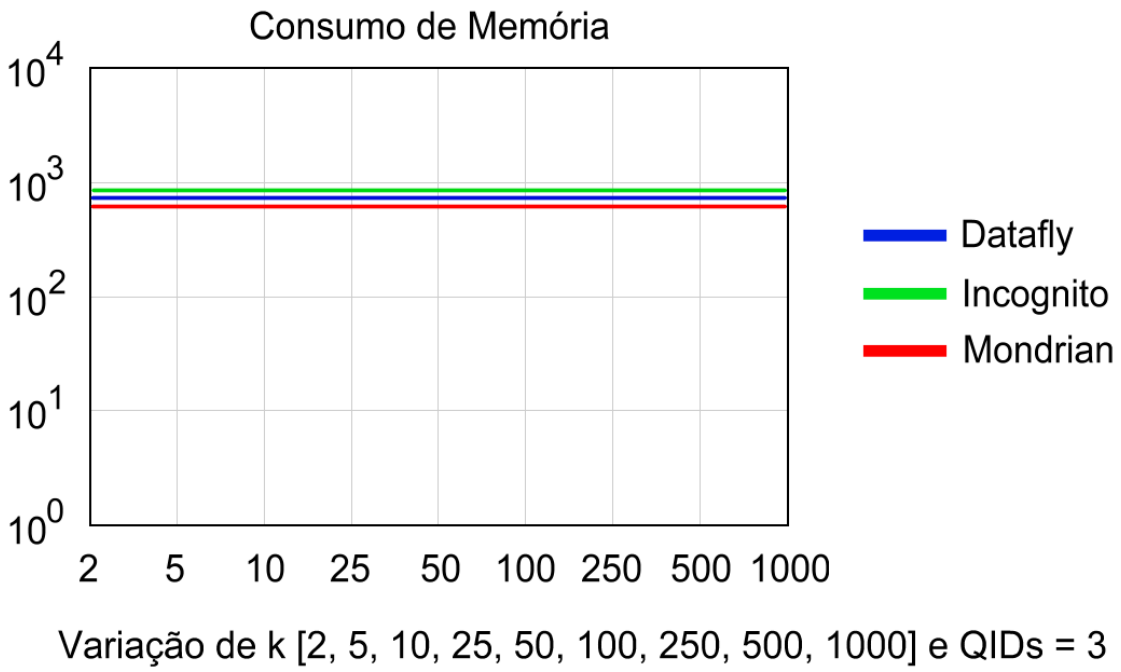
Conforme o valor de k é alterado, as diferenças no número de generalizações ou de nós avaliados não impactam de maneira significativa os algoritmos Incognito e Datafly.

O Mondrian, apresenta um comportamento distinto sobre os outros dois. Ele mostra uma tendência decrescente. Esse comportamento acontece por que à medida que o valor de k aumenta, diminui a quantidade possível de partições que podem satisfazer k. Essa diferença no número de partições mostra a influência que a distribuição de dados dos QIDs tem no desempenho do Mondrian.

4.2.2 Consumo de Memória

O gráfico da Figura 28, mostra o consumo de memória dos 3 algoritmos conforme o número de k aumenta.

Figura 28 - Consumo de memória experimento 2.



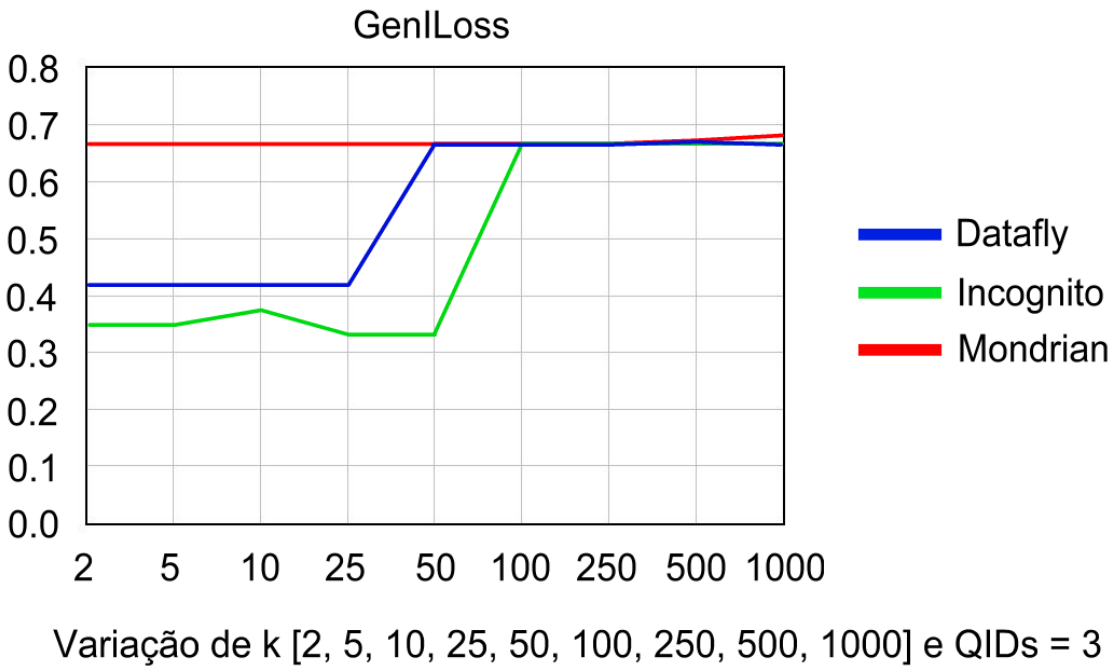
Fonte: (O próprio autor)

Conforme é mostrado no gráfico, o aumento do valor de k não causa impacto no consumo de memória dos algoritmos. O Incognito apresentou um aumento quase imperceptível entre um k e outro. O Datafly não apresentou alterações se mantendo estável em todas as variações do valor de k. O Mondrian apresentou o mesmo comportamento do Datafly. Os algoritmos seguem tendências de serem estáveis quanto ao consumo de memória. Dependendo do tamanho da base de dados utilizada, esse consumo pode apresentar uma variação significativa.

4.2.3 Perda de Informação Generalizada (GenILoss)

O gráfico da figura 29, mostra os resultados da métrica GenILoss para os utilitários de dados conforme aumenta o valor de k.

Figura 29 - GenILoss experimento 2.



Fonte: (O próprio autor)

O Mondrian apresenta o mesmo comportamento de perda de informação à medida que o valor de k vai aumentando. Nesse experimento com o valor de k variando e um QIDs = 3, o comportamento do Mondrian está dentro do esperado, uma vez que ele utiliza apenas um atributo para criar partições e essa métrica retém a perda de precisão nos atributos QID generalizados. Independentemente do valor de k, a pontuação para o GenILoss será a mesma.

Esses fatores fazem com que sua pontuação seja alta. Essa pontuação acontece pelo fato de o Mondrian iniciar a partição dos dados utilizando os estados menos especificos dos seus atributos para aprimora-los posteriormente em cada partição. Se somente um atributo for particionado, os demais retêm o valor mais geral. Isso gera uma penalidade muito grande para esses atributos.

O Datafly apresenta uma variação considerável na pontuação do GenILoss quando o valor de k passa de 25 para 50. Isso acontece por que durante o procedimento de anonimização, ele criou uma solução entre o intervalo {2,25} de k e uma outra quando o valor de k assumiu o valor de 50.

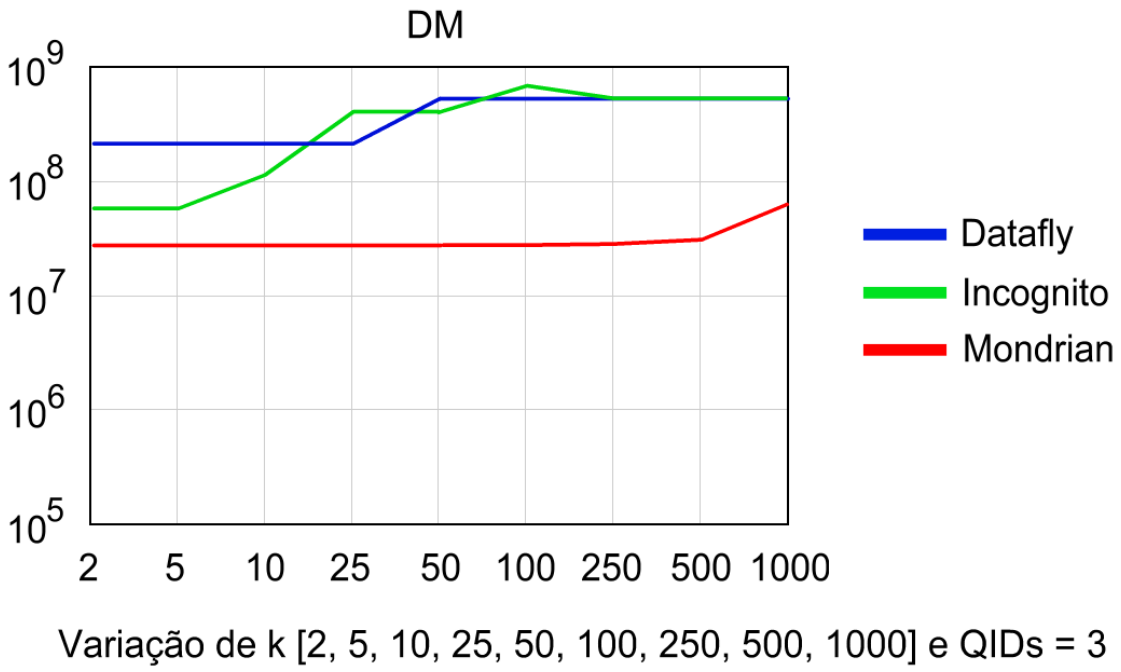
O Incognito apresentou a melhor pontuação para o GenILoss.

Para a configuração de k = 100, todos os algoritmos apresentam a mesma pontuação do GenILoss.

4.2.4 Métrica de Discernibilidade (DM)

O gráfico da figura 30, mostra os resultados da métrica DM para os utilitários de dados conforme aumenta o valor de k.

Figura 30 - DM experimento 2.



Fonte: (O próprio autor)

Observa-se que os 3 algoritmos apresentam comportamentos estáveis apresentando uma sensibilidade baixa conforme o valor de k aumenta.

O Mondrian mostrou ter o melhor desempenho. Conforme o valor de k aumenta, ele cria menos EQs. Isso se deve ao fato de o algoritmo ter como objetivo particionar e maximizar o número de EQs, mas de tamanho maior.

O Incognito apresenta-se como o segundo melhor. Na maioria dos casos, esse algoritmo escolhe uma solução ótima e produz o número máximo de EQs, com exceção dos valores de k {25,100}.

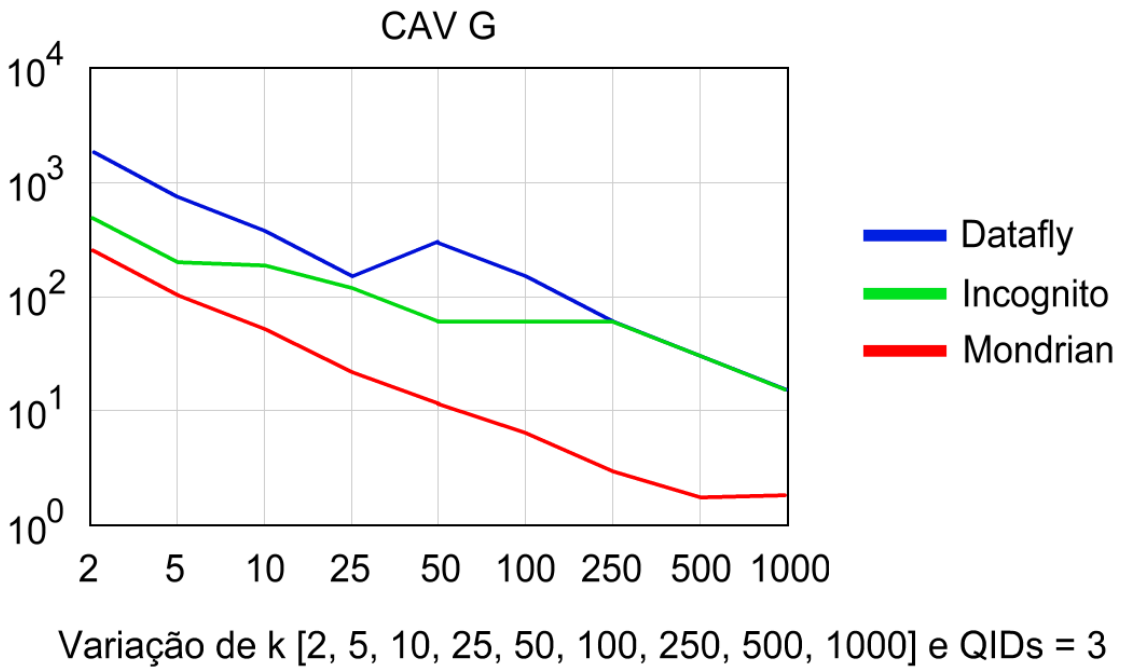
O Datafly apresentou um resultado um pouco melhor que o Incognito. Apesar do Incognito ter produzido mais EQs que o Datafly, um de seus EQs gerados é grande. Isso resultou em um impacto considerável na pontuação do DM do Incognito. Quando k atinge o valor de 25, o Datafly cria 8 EQs onde o maior número de EQ tem 9400 registros. Já o

Incognito, cria 10 EQs, porém o maior possui 18102 registros resultando em uma pontuação geral de DM mais alta.

4.2.5 Tamanho Médio de Classe de Equivalência (CAV G)

O gráfico da Figura 31, mostra os resultados da métrica CAV G para os utilitários de dados conforme aumenta o valor de k.

Figura 31 - CAV G experimento 2.



Fonte: (O próprio autor)

Considerando que as métricas CAV G e DM medem a utilidades dos dados baseado no número de EQs criados, elas tendem a ter o mesmo resultado no desempenho dos 3 algoritmos. Apesar disso, para esse experimento, eles mostraram resultados diferentes. De acordo com os resultados obtidos, as tendências decrescentes mostram que à medida que o valor de k aumenta, o tamanho médio dos EQs gerados ficam bem próximo do cenário ideal onde cada registro é generalizado e agrupado em um EQ composto por k registros, onde o tamanho dos EQs é igual ao k dado.

4.2.6 Resumo do Experimento 2

A diferença do valor de k apresenta pouco ou nenhum impacto direto no desempenho dos algoritmos. Observa-se que o número de partições faz com que o tempo de anonimato diminua. Nas métricas DM e CAV G, que medem o tamanho e o número de EQs criados, o Mondrian supera os outros algoritmos. Para a métrica GenILoss, que captura a transformação dos atributos QID, quando os dados são distorcidos, o Mondrian não consegue executar partições em todos os atributos. Isso gera altas penalidades e faz com que ele tenha o desempenho pior.

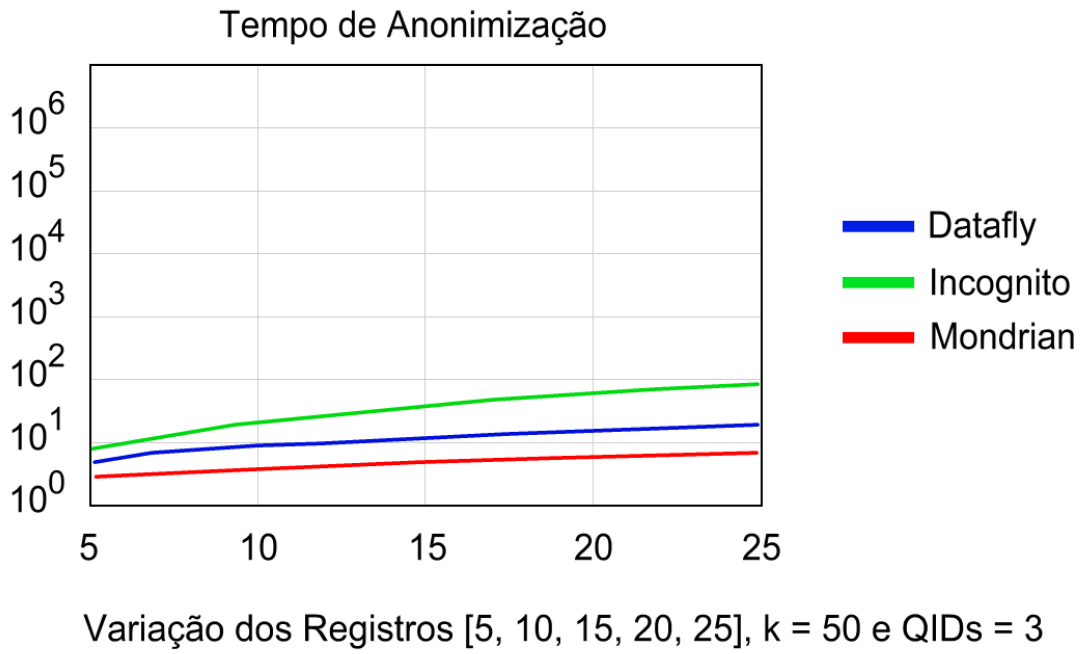
4.3 Experimento 3: Variação no tamanho do arquivo

Para este experimento, serão analisados o tempo de anonimização e o consumo de memória, à medida que o número de registros aumenta. Foram utilizadas as seguintes configurações: Registros = {5 mil, 10 mil e 20 mil}, k = 50 e um | QIDs | = 3.

4.3.1 Tempo de Anonimização

Quanto maior a quantidade de registros para anonimizar, os algoritmos Datafly e Incognito, por tendência, executam uma quantidade menor de generalizações, do mesmo modo que o algoritmo Mondrian tende a executar uma quantidade menor de partições. Apesar de parecer que isso poderia resultar em um menor tempo de execução, durante os testes, foram observados que esse leve aumento de tempo para concluir o processo de anonimização é influenciado diretamente pelo tamanho dos registros. A Figura 32, apresenta esse aumento gradativo de tempo conforme os registros vão crescendo.

Figura 32 - Tempo de anonimização experimento 3.



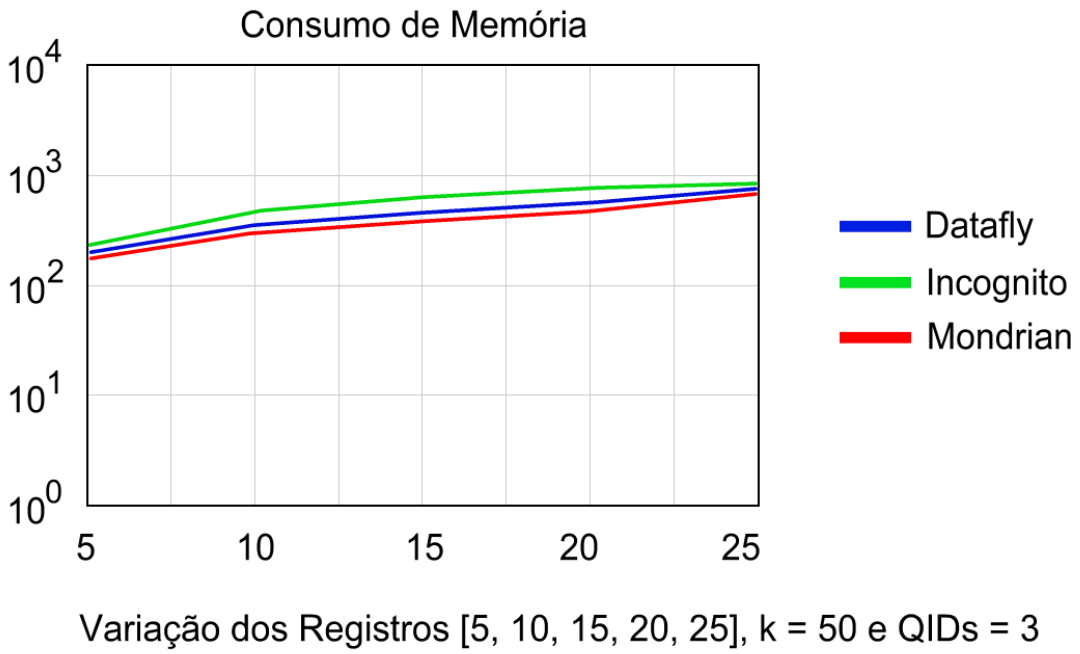
Fonte: (O próprio autor)

4.3.2 Consumo de Memória

Assim como no resultado do tempo gasto para anonimizar os registros, os algoritmos mostraram crescimento no consumo de memória conforme a quantidade de registros aumenta.

No caso do algoritmo Datafly, isso ocorre por que ele necessita de mais memória para verificar a contagem de frequência dos registros e analisar se o valor de k foi satisfeito. Já o Incognito, apresentou um consumo mais alto em relação ao DataFly e ao Mondrian. Isso se deve pelo fato de o Incognito criar a estrutura da rede de generalização para poder fazer a varredura de todos os caminhos possíveis e analisar se esses caminhos, resultam em tabelas anonimizadas para verificar qual delas possui a maior quantidade de classes de equivalência satisfazendo a condição de anonimização do algoritmo. Por fim, o algoritmo Mondrian, assim como o DataFly, também faz a contagem das frequências dos registros, além de calcular a mediana para realizar as partições que ele precisa executar. Apesar do cálculo da mediana e das partições, o Mondrian apresentou um comportamento parecido com o DataFly. A Figura 33 apresenta esses resultados.

Figura 33 - Consumo de memória experimento 3



Fonte: (O próprio autor)

4.3.3 Resumo dos Experimento 3

As variações da quantidade de registros, não apresentaram resultados que afetam diretamente os algoritmos. Cada algoritmo se comportou de acordo com seu procedimento de anonimização. Conclui-se que o tempo gasto e o consumo de memória aumentam de maneira natural, considerando a quantidade de registros que são lidos para executar as generalizações e também as particularidades de cada algoritmo. O algoritmo DataFly foca na profundidade dos VGHs para fazer as generalizações. O Incognito utiliza o tamanho dos VGHs para definir a rede a ser percorrida e o Mondrian utiliza o cálculo de mediana para fazer as partições. Conforme foi mostrado nos resultados do experimento 1 (variação de QIDs), a variação de tempo e consumo de memória só apresentam uma diferença considerada quando se aumenta o número de QIDs. Isso influencia diretamente no comportamento dos algoritmos, ao contrário do valor de k, que apresenta pouca influência.

Capítulo 5 ANÁLISE DE RESULTADOS

Os resultados encontrados nos mostram como a avaliação de várias métricas permite realizar uma análise ampla do desempenho dos algoritmos durante o procedimento de anonimização de dados.

Com os resultados gerados, é possível fornecer algumas informações para ajudar a entender o melhor funcionamento dos algoritmos em cada situação.

Cada algoritmo apresentou um desempenho variado durante os experimentos. Não se pode afirmar que algum algoritmo pode ser considerado melhor do que o outro, pois nenhum deles conseguiu ser melhor em todas as métricas testadas. Um algoritmo que apresentou melhor desempenho para um determinado conjunto de dados, pode ter apresentado uma diferença muito grande com uma outra variação de conjuntos de dados. Dois fatores que impactam diretamente no desempenho dos algoritmos são os requisitos de privacidade e a distribuição dos dados. Os resultados apresentados utilizaram as métricas de eficiência e utilidade dos dados como critério de avaliação.

O algoritmo Incognito, apresentou um crescimento exponencial sobre o tamanho dos QIDs por gastar muito tempo para realizar o processo de anonimização e consumir muita memória. Para VGHs (Hierarquia de generalização de valor) profundos, o espaço para realizar a busca é grande, resultando no tempo gasto para percorrer a rede e analisar em cada estado individual o k-anonimato.

A distribuição de dados e o mecanismo de particionamento afeta diretamente o algoritmo Mondrian quanto ao utilitário de dados. Para distribuições uniformes, o particionamento de mediana é o mais adequado, pois permite que a maioria dos atributos possam ser utilizados para particionar os dados. Se os dados forem distorcidos, não é possível realizar particionamento em todos os seus atributos que afetam a pontuação de perda de informações. Para este caso, o algoritmo apresenta um desempenho ruim. Além disso, valores que não possuem uma ordenação padrão definida, podem ser o motivo de não ser possível realizar o particionamento, pois a ordem em que os valores devem ser processados durante o cálculo de mediana influenciam diretamente na execução do algoritmo. O utilitário de dados é afetado em sua pontuação quando isso acontece. O Mondrian, finaliza sua varredura no

momento em que não forem mais permitidas criar mais partições. Acredita-se que essa estratégia possa beneficiar as métricas do utilitário de dados baseados no tamanho dos EQs, como o DM, pois EQs mais finos são gerados no procedimento de anonimização. No entanto, o DM não é aprimorado para conjuntos de dados que já satisfazem o k-anonimato, por que o número de EQs gerados podem ser menores do que os existentes no conjunto de dados.

O algoritmo Datafly, apresentou o segundo melhor desempenho no tempo de anonimização, consumo de memória e GenILoss. Para as métricas DM e CAV G, ele apresentou o pior desempenho em quase todos os experimentos.

Do ponto de vista das estratégias utilizadas pelos algoritmos, o Mondrian usa generalização baseada em partição, enquanto que o Datafly e o Incognito usam generalização baseada em hierarquia. Essas estratégias possuem dois pontos importantes que precisam ser levadas em conta no ponto de vista do praticante: o tipo de solução que os algoritmos irão produzir para o procedimento de anonimização e os pré-requisitos que precisam ser fornecidos para a aplicação dos algoritmos. O Datafly e o Incognito, possuem parâmetros (hierarquias de generalização) definidos pelo usuário, produzem um anonimato que respeita as restrições definidas pelo VGH. No caso do Mondrian, ele executa um procedimento de anonimização descontrolado. Ao realizar o particionamento, ele cria intervalos de valores dinâmicos ao excluir todas as restrições hierárquicas. Assumindo essa característica, o Mondrian é mais adequado para o procedimento de anonimização de conjunto de dados numéricos (AGHDAM; SONEHARA, 2013). Isso se deve, por que para atributos categóricos, como países por exemplo, qualquer parâmetro de associação aos valores podem comprometer, por exemplo, um agrupamento de países a um determinado continente, pois esses dados não possuem controle de partição sobre os grupos criados. O Datafly e o Incognito, quanto a pré-requisitos de aplicabilidade, necessitam de uma parametrização de um VGH para cada atributo QID. Isso requer conhecimento para oferecer um cenário adequado no VGH. No caso do Mondrian, para executar um particionamento, é necessária uma ação em cada atributo QID.

Do ponto de vista de eficiência, podemos destacar dois termos mencionados: QIDs e nível de anonimato.

- QIDs: Os algoritmos Datafly e Mondrian são melhores nessa situação quando o número de QIDs são grandes. O Incognito peca no desempenho para medidas de eficiência, principalmente em redes grandes de generalização e também onde ocorre valores distintos de atributos.

- Nível de anonimato: O algoritmo Mondrian apresenta melhor desempenho conforme o

valor de k aumenta. De modo geral, como não tem um aumento considerável de tempo, os três algoritmos trabalham bem quando existe a variação do valor de k e mostram um consumo estável de memória.

Tabela 37 – Análise do Tempo de Anonimização

Tempo de Anonimização			
	Experimento 1	Experimento 2	Experimento 3
Incognito	• Crescimento exponencial para VGHs profundos quando aumenta o QIDs. • Em redes grandes o processo de anonimização é descontrolado.	Não apresentam variação significativa de tempo quando aumenta o valor de k.	Crescimento gradual conforme a quantidade de registros vai aumentando. Não gera impacto nos algoritmos.
DataFly	• Possui VGHs definidos pelo usuário. • Produz um anonimato que respeita as restrições definidas pelo VGH, trazendo desempenho mediano conforme aumenta o QID.		
Mondrian	• O mecanismo de particionamento afeta diretamente o algoritmo. • Desempenho mediano para QID maiores. • Em dados distorcidos, não é possível realizar particionamento em todos os seus atributos. • Dados uniformes geram melhores resultados. • Melhor para atributos numéricos.	Apresenta melhor tempo para variações mais altas do valor de K.	

Fonte: (O próprio autor)

Tabela 38 – Análise do consumo de Memória

Consumo de Memória			
	Experimento 1	Experimento 2	Experimento 3
Incognito	Aumenta Exponencialmente com QIDs grandes.	Para o aumento do valor de K, o consumo é equilibrado entre os 3.	Leve variação conforme aumenta a quantidade de registros. Não gera impacto nos algoritmos.
DataFly	Apresenta consumo mediano.		
Mondrian			

Fonte: (O próprio autor)

Do ponto de vista de eficácia, o algoritmo Mondrian apresenta o melhor desempenho quando submetido às métricas DM e CAV G com base no tamanho do grupo. Ele executa uma granularidade mais fina nos EQs melhorando a precisão dos resultados, exceto quando os dados já satisfazem o k-anonimato, pois a utilidade dos dados pode apresentar uma diminuição em relação a essas métricas. O Mondrian é mais indicado quando os dados originais possuem uma distribuição uniforme. Para a métrica GenILoss, que faz a captura da transformação dos atributos, não há um desempenho evidente, apesar que os algoritmos Incognito e Mondrian mostraram bons resultados.

Durante a avaliação dos algoritmos, deve-se levar em consideração os pontos positivos e negativos que serão abordados na avaliação das métricas, para auxiliar na escolha de uma métrica de interesse. Abaixo, alguns pontos relevantes identificados para as métricas utilizadas nos experimentos.

DM: Essa métrica não retém as transformações executadas nos QIDS. Por esse motivo, diversos modelos de anonimização podem apresentar a mesma utilidade de dados quanto ao DM, apesar de que vários QIDs distintos tenham sido anonimizados em cada solução. Alguns registros de EQ podem não ter sidos generalizados, mas mesmo assim sofreram penalidades. No entanto, espera-se que este tipo de métrica capture com precisão a utilidade dos dados para certas aplicações, como resposta de consulta agregada (LEFEVRE et. al, 2006).

Tabela 39 – Análise da (DM) Métrica de Discernibilidade

<table>
<tr><th colspan="3">DM – Métrica de Discernibilidade</th></tr>
<tr><td>Algoritmo</td><td>Experimento 1</td><td>Experimento 2</td></tr>
<tr><td>Incognito</td><td>Desempenho mediano</td><td rowspan="2">Dados que já satisfazem o k-anonimato, recebem valores são altos.</td></tr>
<tr><td>DataFly</td><td>Pior desempenho. Não retém as transformações dos QIDS. Alguns registros podem não ter sidos generalizados, mas mesmo assim sofreram penalidades.</td></tr>
<tr><td>Mondrian</td><td colspan="2">Melhor desempenho, pois executa uma generalização mais fina nos EQs melhorando a precisão dos resultados, exceto se já satisfaz o anonimato, pois pode prejudicar o resultado.</td></tr>
</table>

Fonte: (O próprio autor)

CAV G: Apresenta o mesmo problema da métrica DM além de falhar na distribuição dos registros dentro de cada EQ.

Tabela 40 - Análise do (CAV G) Tamanho Médio de Classe de Equivalência

<table>
<tr><td colspan="3">CAV G</td></tr>
<tr><td>Algoritmo</td><td>Experimento 1</td><td>Experimento 2</td></tr>
<tr><td>Incognito</td><td colspan="2" rowspan="2">Mesma situação do DM. Falha na distribuição dos registros dentro das EQs. Os EQs são levados em consideração nas duas métricas.</td></tr>
<tr><td>DataFly</td></tr>
<tr><td>Mondrian</td><td colspan="2">Melhor desempenho, pois executa uma generalização mais fina nos EQs melhorando a precisão dos resultados, exceto se já satisfaz o anonimato, pois pode prejudicar o resultado.</td></tr>
</table>

Fonte: (O próprio autor)

GenILoss: Essa métrica possui uma grande sensibilidade quando se trata da profundidade VGH dos QIDs. Ao se comparar com outros atributos que possuem um VGH mais raso, um VGH mais profundo que possui um atributo generalizado receberá uma penalidade menor como por exemplo, um atributo do tipo gênero. É importante levar essa característica em conta, pois adicionar certos atributos pode gerar consideravelmente uma degradação do utilitário de dados.

Tabela 41 - Análise do GenILoss - Perda ocorrida ao generalizar um atributo específico

GenILoss		
Algoritmo	Experimento 1	Experimento 2
Incognito	• Sensibilidade no VGH. • Um VGH mais raso receberá uma penalidade maior. • Adicionar certos atributos pode gerar valores altos.	• Só apresentam diferenças consideráveis quando o valor de k é maior que 25. • Afeta o limite geral e específico.
DataFly		
Mondrian	• Valores não numéricos sem padrão, afetam o particionamento. O utilitário de dados é afetado em sua pontuação quando isso acontece.	Apresentou valores iguais em todas as variações de k.

Fonte: (O próprio autor)

Apesar do risco de as métricas serem penalizadas com os pontos fracos mencionados, elas podem ser combinadas para capturar vários aspectos distintos do conjunto dos dados.

Capítulo 6 CONCLUSÃO E TRABALHOS FUTUROS

Foram realizados experimentos para avaliar as características dos três algoritmos de k-anonimização utilizados quanto ao desempenho, termos de eficiência e utilidade dos dados. Para estes experimentos, foram utilizadas implementações públicas disponíveis desses algoritmos. Foram apresentados diversos cenários onde cada um dos algoritmos mostrou bom ou mau desempenho dentro dos termos das métricas submetidas. De acordo com os resultados encontrados, não há como definir um algoritmo de anonimização ótimo de modo geral para todos os cenários aos quais foram analisados, porém, pode-se afirmar que certo algoritmo possui desempenho satisfatório para determinado cenário baseando-se em alguns parâmetros específicos. Na Tabela 42, são apresentadas as indicações mais comuns do uso de cada algoritmo.

Tabela 42 - Indicações dos cenários para cada algoritmo

DataFly	Indicado para arquivos de procedimentos médicos. Necessita de parametrização de VGH para cada atributo QID. Isso requer conhecimento para oferecer um cenário adequado no VGH. Utiliza generalizações únicas. Pode ser utilizado para dados gerais de cadastro.
Incognito	Voltado para uma estrutura de anonimização de grandes conjuntos de dados, ambientes de data center, tanto data centers privados quanto nuvens públicas, e destina-se a ser compatível com estruturas modernas de análise de dados, como conjuntos de dados distribuídos. Utiliza a mesma parametrização do Datafly, porém é mais recomendado para análises mais profundas dos dados, uma vez que a rede de generalização criada por ele, oferece vários caminhos que possibilitam resultados de anonimizações distintos baseados no tamanho do EQs.
Mondrian	Mais utilizado em métodos clássicos de mineração de dados sintáticos de preservação da privacidade, principalmente para base de dados com conteúdo numérico e uniformes.

Fonte: (O próprio autor)

Com os experimentos realizados, foram apresentados diversos fatores que podem ser levados em consideração na escolha de um algoritmo, bem como seus pontos positivos e negativos a partir de um conjunto de métricas de utilidade geral. Os resultados encontrados forneceram várias evidências da dificuldade de selecionar um algoritmo de anonimização, apresentando a real necessidade de desenvolver novas metodologias que possam auxiliar os profissionais no processo de identificação dos algoritmos de anonimização para definir qual algoritmo é mais adequado para uma determinada demanda com base em diversas diretrizes que podem facilitar a adoção de novas técnicas para a preservação de privacidade.

6.1 Trabalhos Futuros

Para trabalhos futuros, é fundamental identificar novas formas de medir o comportamento dos algoritmos e criar novas métricas para padronizar a comparação dos algoritmos levando em conta as necessidades de cada usuário. Submeter os resultados encontrados a um processo de re-identificação dos dados. Analisar esses dados anonimizados utilizando métricas voltadas para re-identificação. Verificar a possibilidade de comparar o

modelo k-anonymity com outros modelos de privacidade existentes, como por exemplo, l-diversity ou t-closeness explorando os parâmetros que diferenciam um dos outros.

REFERÊNCIAS

AGHDAM, M. R. S.; SONEHARA, N. On enhancing data utility in k-anonymization for data without hierarchical taxonomies. **International Journal of Cyber-Security and Digital Forensics**, Freienbach, v. 2, n. 2, p. 12-23, 2013. Disponível em: https://oaji.net/articles/2014/541-1394065059.pdf. Acesso em: 17 abr. 2023.

AYALA-RIVERA, V. *et al*. A systematic comparison and evaluation of k-anonymization algorithms for practitioners. **Transactions on Data Privacy**, Skövde, v. 7, n. 3, p. 337-370, 2014. Disponível em: http://www.tdp.cat/issues11/tdp.a169a14.pdf. Acesso em: 17 abr. 2023.

BASSO, T. *et al*. Challenges on anonymity, privacy, and big data. *In*: LATIN-AMERICAN SYMPOSIUM ON DEPENDABLE COMPUTING (LADC), 7., 2016, Cali. **Proceedings** [...]. Piscataway: IEEE, 2016. p. 164-171. DOI: https://doi.org/10.1109/LADC.2016.34.

BAYARDO, R. J.; AGRAWAL, R. Data privacy through optimal k-anonymization. *In*: INTERNATIONAL CONFERENCE ON DATA ENGINEERING (ICDE'05), 21., 2005, Tokyo. **Proceedings** [...]. Piscataway: IEEE, 2005. p. 217-228. DOI: https://doi.org/10.1109/ICDE.2005.42.

BBC. Personnel records stolen from Mod. *In*: BBC News Channel. London, 27 Sept. 2008. Disponível em: http://news.bbc.co.uk/2/hi/uk_news/england/gloucestershire/7639006.stm. Acesso em: 17 abr. 2023.

BRANCO JUNIOR, E.; MACHADO, J.; MONTEIRO, J. Estratégias para proteção da privacidade de dados armazenados na nuvem. *In*: L'OSCIO, B. F.; HARA, C. S.; MARTINS, V. (org.). **Tópicos em gerenciamento de dados e informações**. Curitiba: Sociedade Brasileira de Computação, 2014. p. 46-75.

BRASIL. Governo Federal. **Dados Abertos**. Brasília, DF, 2021. Disponível em: https://dados.gov.br/home. Acesso em: 13 abr. 2021.

BRASIL. **Lei nº 13.709, de 14 de agosto de 2018**. Lei Geral de Proteção de Dados Pessoais (LGPD). Brasília, DF: Presidência da República, 2018. Disponível em: http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2018/Lei/L13709.htm. Acesso em: 04 mar. 2022. Acesso em: 17 abr. 2023.

BRASIL. **Lei nº 13.853, de 8 de julho de 2019**. Altera a Lei nº 13.709, de 14 de agosto de 2018, para dispor sobre a proteção de dados pessoais e para criar a Autoridade Nacional de Proteção de Dados; e dá outras providências. Brasília, DF: Presidência da República, 2019. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2019/lei/l13853.htm. Acesso em: 17 abr. 2023.

CAMENISCH, J.; FISCHER-HÜBNER, S.; RANNENBERG, K. (ed.). **Privacy and identity management for life**. [*S. l.*]: Springer, 2011. Disponível em: http://dblp.uni-trier.de/db/books/collections/primelife2011.html. Acesso em: 17 abr. 2023.

CORREIOS. **Tudo sobre CEP**. Brasília, DF, 2020. Disponível em: https://www.correios.com.br/enviar/precisa-de-ajuda/tudo-sobre-cep. Acesso em: 17 abr. 2023.

EL EMAM, K.; DANKAR, F. K. Protecting privacy using k-anonymity. **Journal of the American Medical Informatics Association**, Philadelphia, v. 15, n. 5, p. 627-637, 2008. DOI: https://doi.org/10.1197/jamia.M2716.

GANTZ, J.; REINSEL, D. The digital universe in 2020: big data, bigger digital shadows, and biggest growth in the far east. **IDC Analyze the Future**, Framingham, 2012. Disponível em: https://datastorageasean.com/sites/default/files/idc-the-digital-universe-in-2020.pdf. Acesso em: 17 abr. 2023.

GOODIN, D. Poorly anonymized logs reveal NYC cab drivers' detailed whereabouts. *In*: ARS Technica. Boston, 23 June 2014. Disponível em: https://arstechnica.com/tech-policy/2014/06/poorly-anonymized-logs-reveal-nyc-cab-drivers-detailed-whereabouts/. Acesso em: 17 abr. 2023.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). **Censo 2010**. Brasília, DF: IBGE, 2010. Disponível em: https://censo2010.ibge.gov.br. Acesso em: 17 abr. 2023.

IYENGAR, V. S. **Transforming data to satisfy privacy constraints**. [*S. l.*]: ACM, 2002. p. 279-288. Disponível em: http://dblp.uni-trier.de/db/conf/kdd/kdd2002.html#Iyengar02. Acesso em: 17 abr. 2023.

KANTARCIOGLU, M. Anonymization ioolbox. *In*: UTD Anonymization Toolbox. Richardson, 2019. Disponível em: https://cs.utdallas.edu/dspl/cgi-bin/toolbox/. Acesso em: 17 abr. 2023.

KOHAVI, R.; BECKER, B. Adult data set. *In*: UCI Machien Learning Repository. Irvine, 1996. Disponível em: https://archive.ics.uci.edu/ml/datasets/Adult. Acesso em: 17 abr. 2023.

LEFEVRE, K.; DEWITT, D. J.; RAMAKRISHNAN, R. Incognito: efficient full-domain k-anonymity. *In*: ACM SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, 2005, Baltimore. **Proceedings** [...]. New York: Association for Computing Machinery, 2005. p. 49-60. DOI: https://doi.org/10.1145/1066157.1066164.

LEFEVRE, K.; DEWITT, D. J.; RAMAKRISHNAN, R. Mondrian: multidimensional k-anonymity. *In*: INTERNATIONAL CONFERENCE ON DATA ENGINEERING (ICDE'06), 22., 2006, Atlanta. **Proceedings** [...]. Piscataway: IEEE, 2006. p. 25-25. DOI: https://doi.org/10.1109/ICDE.2006.101.

MCCANN, E. Walgreens company announces data breach. *In*: HEALTH Care IT News. [*S. l.*], 25 Feb. 2013. Disponível em: https://www.healthcareitnews.com/news/walgreens-company-announces-data-breach. Acesso em: 17 abr. 2023.

NARAYANAN, A.; SHMATIKOV, V. Privacy and security myths and fallacies of "personally identifiable information". **Communications of the ACM**, New York, v. 53, n. 6, p. 24-26, 2010. DOI: https://doi.org/10.1145/1743546.1743558.

NARAYANAN, A.; SHMATIKOV, V. Robust de-anonymization of large sparse datasets. *In*: IEEE SYMPOSIUM ON SECURITY AND PRIVACY (SP 2008), 2008, Oakland. **Proceedings** [...]. Piscataway: IEEE, 2008. p. 111-125. DOI: https://doi.org/10.1109/SP.2008.33.

NERGIZ, M. E.; CLIFTON, C. Thoughts on k-anonymization. **Data & Knowledge Engineering**, [*s. l.*], v. 63, n. 3, p. 622-645, 2007. DOI: https://doi.org/10.1016/j.datak.2007.03.009.

OHM, P. Broken promises of privacy: responding to the surprising failure of anonymization. **UCLA Law Review**, Los Angeles, v. 57, p. 1701-1777, 2010. Disponível em: https://www.uclalawreview.org/pdf/57-6-3.pdf. Acesso em: 17 abr. 2023.

PANDURANGAN, V. On taxis and rainbows: lessons from NYC's improperly anonymized taxi logs. *In*: MEDIUM. [*S. l.*], 21 June 2014. Disponível em: https://tech.vijayp.ca/of-taxis-and-rainbows-f6bc289679a1. Acesso em: 17 abr. 2023.

PATEL, T.; AMIN, K. A study on k-anonymity, l-diversity, and t-closeness: techniques of privacy preservation data publishing. **IJIRST International Journal for Innovative Research in Science & Technology**, Gujarat, v. 6, n. 6, p. 19-24, 2019. Disponível em: http://www.ijirst.org/articles/IJIRSTV6I6015.pdf. Acesso em: 17 abr. 2023.

PONEMON INSTITUTE. **2018 cost of data breach study**: impact of business continuity management. Traverse City: Ponemon Institute, 2018. Disponível em: https://www.ibm.com/downloads/cas/4DNXZYWK. Acesso em: 17 abr. 2023.

QUICK, M. *et al*. World's biggest data breaches and hacks. *In*: INFORMATION is Beautiful. [*S. l.*], 2018. Disponível em: http://www. informationisbeautiful.net/visualizations/worlds- biggest-data- breacheshacks/. Acesso em: 4 mar. 2022.

SAMARATI, P. Protecting respondents identities in microdata release. **IEEE Transactions on Knowledge and Data Engineering**, Piscataway, v. 13, n. 6, p. 1010-1027, 2001. DOI: https://doi.org/10.1109/69.971193.

SAMARATI, P.; SWEENEY, L. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. *In*: IEEE SYMPOSIUM ON RESEARCH IN SECURITY AND PRIVACY (S&P), 1998, Oakland. **Proceedings** [...]. Piscataway: IEEE, 1998. Disponível em: https://dataprivacylab.org/dataprivacy/projects/kanonymity/paper3.pdf. Acesso em: 17 abr. 2023.

SWEENEY, L. Achieving k-anonymity privacy protection using generalization and suppression. **International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems**, Singapore, v. 10, n. 5, p. 571-588, 2002. DOI: https://doi.org/10.1142/S021848850200165X.

THE LOCAL. **Swedish authority handed over 'keys to the Kingdom' in IT security slip-up**. Stockholm, 17 July 2017. Disponível em: https://www.thelocal.se/20170717/swedish-authorityhanded-over-keys-to-the-kingdom-in-it-security-slip-up. Acesso em: 17 abr. 2023.