TÉCNICAS DE IMPLEMENTAÇÃO DE ANONIMIZAÇÃO DE DADOS

Luiz Eduardo da Cunha Andrade[1], Lucas Baggio Figueira[2]

[1],[2]Faculdade de Tecnologia de Ribeirão Preto (FATEC)

Ribeirão Preto, SP – Brasil

[1]luiz.andrade8@fatec.sp.gov.br,

[2]lucas.figueira@fatec.sp.gov.br

Resumo. *Com o avanço das tecnologias de comunicação e aumento do tráfego de dados no cenário mundial, avança também o crescente número de fraudes e vazamentos de dados, tanto de pequenas empresas quanto de grandes corporações e conglomerados. Diante disto, se faz cada vez mais necessário o uso de técnicas e metodologias para aumentar as camadas de segurança do processamento, tráfego e armazenamento de dados. Este artigo apresenta técnicas para anonimização de dados pessoais, passíveis de fraude ou corrupção, com o intuito de preservar a integridade e a segurança dos dados. A pesquisa mostra a importância e aplicação de algumas destas técnicas.*

Abstract. *With the advancement of communication technologies and the increase in data traffic on the world stage, too the number of frauds and data leaks, from small companies to large corporations and conglomerates. Exposing this, it is increasingly necessary to use techniques and methodologies to increase the security layers of processing, traffic and data storage. This article presents techniques for anonymizing personal data, susceptible to fraud or corruption, in order to preserve data integrity and security. Research shows the importance and application of some of these techniques.*

1. Introdução

Em 2021, vivemos uma era de criação e compartilhamento de informações e dados sem precedentes e nunca antes visto. Estima-se que a humanidade produza cerca de 1,7 megabytes por segundo atualmente. Isso equivale a 2,5 quintilhões bytes de dados por dia (BULAO, 2021).

Diante de tamanho fluxo de movimentação de dados, comuns são notícias de ataques cibernéticos, realizados por hackers que exigem resgate pelos dados, bem como o vazamento de dados, por vezes ocorridos não somente por ataques, mas por falhas de segurança que permitiram o vazamento de dados através de métodos não tão avançados, como o vazamento ocorrido no segundo semestre de 2020, quando ocorreu o vazamento de credenciais de acesso ao sistema de notificações de Covid-19, do Ministério da Saúde do Governo Brasileiro, em que um arquivo oculto contendo as credenciais de acesso estava acessível no código-fonte do site da plataforma e que, através destas credenciais, concediam acesso ao banco de dados com informações pessoais de milhões de pessoas(FEITOSA Jr, 2020).

Em 2021, ao menos oito grandes vazamentos de dados aconteceram no Brasil (COUTINHO, 2021). Diante do exposto, se faz necessário focar em camadas de segurança cada vez mais robustas na manipulação e tráfego de dados. Dentro deste cenário, o presente trabalho tem por objetivo analisar técnicas de anonimação de dados, que consiste em tornar o dado anônimo, em relação a personificação da informação contida no dado(BIONI, 2020), com o intuito de que, diante de um vazamento de dados, a informação contida se torne inutilizável devido às informações possuírem um espectro anônimo, sem sua correta identificação ou valor.

Este processo é realizado aplicando-se uma ou mais técnicas de anonimação e verificando o risco de re-identificação. Todavia, é cada vez mais recorrente a publicação de estudos que demonstram ser o processo de anonimização algo falível. A representação simbólica de que os vínculos de identificação de uma base de dados poderiam ser completamente eliminados, garantindo-se, com 100% (cem por cento) de eficiência, o anonimato das pessoas, é um mito (NARAYANAN; SHMATIKOV, 2010).

Contudo, a aplicação de técnicas de anonimação de dados se faz necessária, uma vez que é um requisito de proteção dos dados pessoais exigida pela Lei Geral de Proteção de Dados (LGPD) que prevê, em seu artigo 18, inciso IV, que o titular dos dados pessoais tem direito, entre outros, à anonimização, bloqueio ou eliminação de dados desnecessários, excessivos ou tratados em desconformidade com o disposto na Lei (MONTANHA, 2020).

Deste modo, a anonimização de dados passou a ser um direito do titular de dados pessoais e uma obrigação daquele que armazena tais dados.

2. Regulamentação

Diante do cenário citado, de reincidentes vazamentos de dados pessoais por grandes empresas e governos, diversos países implementaram legislações específicas para tratar e combater os crimes virtuais relacionados a vazamento e utilização de dados de terceiros sem prévio consentimento. É possível citar a General Data Protection Regulation (GDPR), que consiste em um regulamento do Parlamento Europeu e Conselho da União Europeia, que define regras sobre a privacidade e proteção de dados de cidadãos da União Europeia, vigente desde maio de 2018 (SOUZA, 2018). O EUA (Estados Unidos da América) não possui uma única lei, mas sim, diversas leis promulgadas em níveis federal e estadual, que regulamentam a proteção dos dados, como a Driver's Privacy Protection Act (DPPA), que define regras e cuidados dos departamentos estaduais de veículos a motor quanto às informações pessoais da população, e a New York Stop Hacks and Improve Electronic Data Security Act (NY SHIELD), lei do estado de Nova York, que determina uma maior transparência e cuidado com que determinadas empresas devem ter ao trabalhar com dados pessoais (GATEFY, 2021). No Brasil, existe a Lei Geral de Proteção de Dados (LGPD), vigente desde setembro de 2020, que também define parâmetros para o armazenamento e utilização de dados pessoais de terceiros, que na lei em questão é chamado de tratamento, com definição de multas para as empresas que não se adequem as diretrizes previstas na lei (NONES, 2021).

Além das leis citadas, diversos outros países estão elaborando e implementando leis para definir limites e responsabilidades no uso de dados, o que embasa que, além das necessidades de segurança para dispor de credibilidade e confiabilidade na prestação de serviços digitais pelas empresas, já que um vazamento de dados prejudicaria muito a visibilidade de uma empresa em qualquer lugar no mundo, se faz necessário atender os requisitos jurídicos impostos por essas leis.

3. Métodos e Técnicas

Para aplicar as técnicas de anonimação de dados sobre uma base de dados, antes de qualquer etapa é preciso analisar o contexto dos dados e informações que sofrerão a transformação para assim definir a melhor técnica e abordagem. A depender das características dos dados, as técnicas analisadas podem não fazer sentido ou não promover o efeito esperado (IMPERVA, 2021).

Também se faz necessário analisar a natureza dos dados: numérico ou categórico. Um dado numérico é constituído por um valor numérico, enquanto um dado categórico é aquele que pode ser visto como um grupo distinto de alguma propriedade qualitativa (RENZE, 2019) . Em ambos os casos, a análise do conteúdo do atributo, o contexto da utilidade da informação contida nos dados e o risco de re-identificação é que irão definir a melhor técnica a ser adotada.

3.1. Generalização (Generalization)

Esta técnica consiste na redução da precisão dos dados. Os valores dos atributos são alterados por outros semanticamente semelhantes, porém, menos específicos, preservando assim a veracidade dos dados (PDPC, 2018).

Pode ser aplicada tanto em atributos categóricos como também em atributos numéricos. Ao aplicar em atributos numéricos, é possível atribuir um intervalo. Deste modo, o dado permanece útil do ponto de vista analítico, sem deixar explícito qual seria a informação real. Em atributos categóricos a estratégia é semelhante, contudo, com uma hierarquia semântica que abstrai o valor exato do dado, mas preservando o sentido(IMPERVA, 2020).

Generalização

user_id	Idade	Endereço
259	18	Av. Carlos Alberto, 1292, 57400-654
124	21	Rua Tocantins, 200, 74324-259
634	38	Rua Goiás, 32, 27477-300
87	19	Av. João Roberto, 387, 37950-600
287	26	Rua Maria Joaquina, 89, 09865-367
561	31	Av. Minas Gerais, 2784, 14784-746

**Tabela antes da aplicação da técnica*

user_id	Idade	Endereço
259	10-20	Av. Carlos Alberto
124	20-30	Rua Tocantins
634	30-40	Rua Goiás
87	10-20	Av. João Roberto
287	20-30	Rua Maria Joaquina
561	30-40	Av. Minas Gerais

**Tabela depois da aplicação da técnica*

Figura 1. Técnica "Generalização"
Fonte: Próprio Autor, 2021

Na Figura 1 é possível observar a aplicação da técnica nas colunas "Idade" e "Endereço". Na coluna "Idade", ocorreu a transformação do valor da idade do indivíduo em um intervalo que contempla o valor real da idade no dado original. Já na coluna "Endereço" houve uma abstração hierárquica do valor do campo, substituindo o endereço completo por apenas o logradouro. Essas generalizações possibilitam preservar a informação em um determinado nível de poder analítico e ocultam o seu valor exato.

3.2. Agregação de dados (Data aggregation)

Esta técnica consiste na conversão de um conjunto de dados em uma lista de valores resumidos (PDPC, 2018). Em uma outra interpretação, ao invés de uma tabela com diversas entradas contendo dados pessoais, cria-se novas colunas que preservam as propriedades estatísticas da base de dados e mascaram a identidade dos indivíduos portadores das informações.

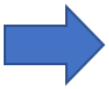
Diferencia-se a técnica de Agregação de dados em relação a técnica de Generalização por abstrair o valor de cada célula de uma determinada coluna, enquanto que a Generalização transforma a estrutura do conjunto de dados, removendo colunas e criando novas.

Nesta técnica é importante ter cuidado com o tamanho dos grupos agregados para evitar que possuam poucas entradas. Para um ataque com informações suficientes, um grupo com um único indivíduo pode conter informações suficientes para a re-identificação.

Agregação

Doador	Renda	Valor Doado
Doador A	3000	150
Doador B	2600	400
Doador C	2800	120
Doador D	6500	500
Doador E	5200	470
Doador F	4700	350

**Tabela antes da aplicação da técnica*



Renda Mensal	Nº de doações	Total Doado
2000 - 2999	2	520
3000 - 3999	1	150
4000 - 4999	1	350
5000 - 7000	2	970

**Tabela depois da aplicação da técnica*

Figura 2. Técnica "Agregação"
Fonte: Próprio Autor, 2021

Na Figura 2 é possível observar que a informação "Doador" foi ocultada através da agregação em intervalos dos valores de salário mensal, assim como a informação "Total Doado" deixou de identificar um único indivíduo para corresponder à soma de cada conjunto de intervalos.

3.3. Mascaramento de dados (Data masking)

Esta técnica, também conhecida como mascaramento de caracteres, tem como objetivo substituir caracteres do valor de uma coluna ou atributo por símbolos como o asterisco ou o caracter "X" (IMPERVA, 2020).

Normalmente o mascaramento é feito apenas em parte do valor do atributo, deixando apenas parte do valor oculto. Desta forma, a técnica pode ser aplicada em uma quantidade fixa de caracteres, como por exemplo ocorre com cartões de crédito que, por possuírem um tamanho e comprimento padrão, muito comumente são apresentados com apenas os últimos quatro números visíveis, tendo o restante da numeração ocultada a fim de evitar fraudes por uso indevido. Ao utilizar a técnica de mascaramento de dados é preciso analisar qual parte e tamanho da informação são mais apropriados a serem ocultados, de forma que a parte visível da informação não permita a re-identificação do dado(PDPC, 2018).

Para o caso especial em que o proprietário do dado deve ser capaz de identificá-lo, como o mascaramento dos últimos dígitos de um CPF na relação de candidatos do resultado de um concurso ou avaliação, foge-se do propósito das técnicas de anonimação, que é inviabilizar todas as possibilidade de reconhecimento da informação por qualquer indivíduo. Desta forma, esse cenário deve ser tratado fora do escopo da tarefa de impersonificação.

Mascaramento

Matrícula	Plano	Email
12098872077	Básico	joao.santos@mail.com
75262426852	Básico	jose.silva@mail.com
98018545780	Premium	maria.coelho@mail.com
84578521457	Premium	eduardo.souza@mail.com
24578412585	Ultra	silvia.lima@mail.com
45715236987	Ultra	luana.martins@mail.com

**Tabela antes da aplicação da técnica*

Matrícula	Plano	Email
120988*****	Básico	joao.sa****@mail.com
752624*****	Básico	jose.s****@mail.com
980185*****	Premium	maria.co****@mail.com
845785*****	Premium	eduardo.s****@mail.com
245784*****	Ultra	silvia.****@mail.com
457152*****	Ultra	luana.mar****@mail.com

**Tabela depois da aplicação da técnica*

Figura 3. Técnica "Mascaramento"
Fonte: Próprio Autor, 2021

Na Figura 3 é possível observar que a técnica foi aplicada em duas colunas com aplicações distintas. Tanto na coluna "Matrícula" quanto na coluna "Email" foram ocultadas uma quantidade fixa de caracteres, porém, em posições diferentes, mas com o mesmo intuito.

3.4. Pseudo Anonimação

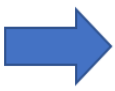
Esta técnica, também identificada como Pseudoanonimação, consiste na substituição de um identificador com valores falsos. Esses dados falsos devem ser únicos e não apresentar qualquer relação com os dados originais (PDPC, 2018).

Os pseudônimos podem ser aleatórios ou pré-definidos, mas, de toda forma, a informação do dado original é completamente perdida, o que acarreta em uma grande perda de utilidade das informações. Uma possibilidade de preservar o dado original é utilizar pseudônimos persistentes, ou seja, usar o mesmo pseudônimo para identificar o mesmo indivíduo em bases de dados diferentes.

Pseudoanonimação

Aluno	Nota	Horas de estudo
Pedro Silva	10,0	55
Maria Eduarda	7,0	42
Lucas Lima	9,0	40
Carlos Melo	8,5	50
Beatriz Gomes	6,0	30
Jaqueline Silva	7,5	32

**Tabela antes da aplicação da técnica*



Aluno	Nota	Horas de estudo
12567	10,0	55
34875	7,0	42
21478	9,0	40
48007	8,5	50
36902	6,0	30
18456	7,5	32

**Tabela depois da aplicação da técnica*

Figura 4. Técnica "Pseudoanonimação"
Fonte: Próprio Autor, 2021

Na Figura 4 é possível observar que a técnica foi aplicada na coluna "Aluno", substituindo a identificação do nome dos indivíduos por um valor numérico possivelmente aleatório e não

sequencial.

Como esta técnica provoca a perda total do dado original, comumente é utilizada uma outra tabela que relaciona os dados antes da anonimação com os dados após o processo, e é preciso que essa tabela referencial seja preservada em ambiente seguro.

3.5. Perturbação de dados (Data perturbation)

Esta técnica consiste na substituição dos valores originais dos atributos por outros valores levemente distintos e diferentes dos originais (PDPC, 2018). Diferente da técnica de Generalização, esta abordagem não preserva a veracidade dos dados originais, por isso, a utilidade do dado é fortemente afetada, mesmo quando a modificação do valor é sutil.

Nesta técnica, o nível de perturbação dos dados deve ser proporcional ao intervalo dos valores do atributo. Desta forma, se a base de dados for pequena, o efeito será menor. Para bases de dados grandes, a diferença e impacto nos valores dos dados poderá ser maior. A perturbação é realizada através da adição de ruído, tipicamente em atributos numéricos, que consiste na adição ou multiplicação do valor original por um valor diferente pré definido.

Desta forma, determinadas propriedades estatísticas são preservadas, como média e correlação. Por outro lado, é possível produzir valores sem significado expressivo.

Perturbação

Pessoa	Altura(cm)	Peso(kg)	Idade
12567	160	50	30
34875	177	70	36
21478	158	46	20
48007	173	75	22
36902	169	82	44

**Tabela antes da aplicação da técnica*

Pessoa	Altura(cm)	Peso(kg)	Idade
12567	160	51	30
34875	175	69	36
21478	160	45	18
48007	175	75	21
36902	170	81	42

**Tabela depois da aplicação da técnica*

Figura 5. Técnica "Perturbação"
Fonte: Próprio Autor, 2021

Na Figura 5 é possível observar que a técnica foi aplicada e que, por consequência, os dados sofreram uma pequena variação quando comparados aos dados originais. Neste cenário foi aplicado um arredondamento para os múltiplos mais próximos de cinco, na coluna "Altura", e de três, nas colunas "Peso" e "Idades".

4. Conclusão

Devido a diversas novas legislações que surgem em todos os países do mundo, se faz necessário para todo profissional que tem contato com dados pessoais, o conhecimento das aplicações e consequências das técnicas de anonimação . Novas técnicas surgem com regularidade e é importante o empenho da comunidade para criar métodos e processos que garantam a maior cobertura possível de camadas de segurança que não permitam a identificação, associação e mal uso

de dados vazados.

No que diz respeito à aplicação das técnicas de anonimação de dados, o contexto e uso dos dados, bem como o risco de re-identificação, são os fatores essenciais para a definição de quais atributos são mais favoráveis para a impersonificação.

Definidos os atributos a sofrerem o processo de anonimação dos dados, é possível então definir a melhor técnica e abordagem, com a finalidade de preservar ao máximo a utilidade dos dados sem comprometer o mascaramento da informação.Desta forma, é imprescindível que exista um conhecimento aprofundado do comportamento de cada estratégia, para que seja possível evitar que o dado seja associado ao seu proprietário, mas também para que o dado não perca o seu valor analítico.

As técnicas de anonimação de dados não se limitam às técnicas apresentadas neste artigo, sendo o foco das informações contidas neste artigo apresentar as possibilidades, técnicas e processos já existentes, bem como abrir possibilidades de novas técnicas e uso de tais informações.

5. Referências

.

BULAO, Jacquelyn. How Much Data Is Created Every Day in 2021?. Disponível em: <http://www.ufrgs.br/niee/eventos/CIIEE/2002/programacao/Demonstracoes.pdf>. Acesso em: 1 out 2021.

COUTINHO, Dimítria. Ao menos oito vazamentos de dados aconteceram no Brasil em 2021; quem é punido?. Disponível em: <https://tecnologia.ig.com.br/2021-03-28/ao-menos-oito-vazamentos-de-dados-aconteceram-no-brasil-em-2021--quem-e-punido-.html>. Acesso em: 1 out 2021.

NARAYANAN, Arvind; SHMATIKOV, Vitaly. Privacy and Security Myths and Fallacies of “Personally Identifiable Information”. Disponível em: <https://www.cs.utexas.edu/~shmat/shmat_cacm10.pdf>. Acesso em: 1 out 2021.

SOUZA, Ramon De. Mas, afinal, o que é a lei GDPR e como ela afeta os brasileiros?. Disponível em: <https://canaltech.com.br/legislacao/mas-afinal-o-que-e-a-lei-gdpr-e-como-ela-afeta-os-brasileiros-114370/>. Acesso em: 1 out 2021.

NONES, Fernada. LGPD: o que diz a lei brasileira de proteção de dados e como ela pode impactar a estratégia de marketing de sua empresa. Disponível em: <https://resultadosdigitais.com.br/blog/o-que-e-lgpd/>. Acesso em: 1 out 2021.

BIONI, Bruno. Compreendendo o conceito de anonimização e dado anonimizado.. Disponível em: <http://genjuridico.com.br/2020/08/05/conceito-anonimizacao-dado-anonimizado>. Acesso em: 1 out 2021.

BRITO, Felipe Timbó. MACHADO, Javam. Preservação de Privacidade de Dados: Fundamentos, Técnicas e Aplicações. Disponível em: <https://www.researchgate.net/publication/318726149_Preservacao_de_Privacidade_de_Dados_Fundamentos_Tecnicas_e_Aplicacoes>. Acesso em: 1 out 2021.

CORPORATE FINANCE INSTITUTE. Data Anonymization: The process of preserving private or confidential information by deleting or encoding identifiers that link individuals and the stored data. Disponível em: <https://corporatefinanceinstitute.com/resources/knowledge/other/data-anonymization/>. Acesso em: 1 out 2021.

IMPERVA. What is Data Anonymization. Disponível em: <https://www.imperva.com/learn/data-security/anonymization/>. Acesso em: 1 out 2021.

GATEFY. Como funcionam as leis de proteção de dados nos Estados Unidos. Disponível em: <Como são as leis de proteção de dados nos Estados Unidos - Gatefy>. Acesso em: 1 out 2021.

PDPC (PERSONAL DATA PROTECTION COMISSION SINGAPURE). Guide to Basic Data Anonymisation Techniques – 2018. Disponível em: <https://iapp.org/media/pdf/resource_center/Guide_to_Anonymisation.pdf>. Acesso em: 1 out 2021.

VASCONCELLOS, Hygino. Vazamento de dados de 220 milhões de pessoas: o que sabemos e quão grave é. Disponível em: <https://www.uol.com.br/tilt/noticias/redacao/2021/01/28/vazamento-expoe-dados-de-220-mi-de-brasileiros-origem-pode-ser-cruzada.htm>. Acesso em: 1 out 2021.

FEITOSA Jr, Alessandro. Senhas do Ministério da Saúde para sistema de notificação de Covid-19 também ficaram expostas em junho, diz ONG. Disponível em: <https://g1.globo.com/economia/tecnologia/noticia/2020/11/27/senhas-do-ministerio-da-saude-para-sistema-de-notificacao-de-covid-19-tambem-ficaram-expostas-em-junho-diz-ong.ghtml>. Acesso em: 13 nov 2021.

MONTANHA, Aleksandro. LGPD: a importância da anonimização de dados. Disponível em: <https://opadvogados.com/lgpd-a-importancia-da-anonimizacao-de-dados/>. Acesso em: 13 nov 2021.

NEGRI, Sérgio Marcos Carvalho de Ávila. GIOVANINI, Carolina Fiorini Ramos. Dados não pessoais: a retórica da anonimização no enfrentamento à covid-19 e o privacywashing. Disponível em: <https://revista.internetlab.org.br/dados-nao-pessoais-a-retorica-da-anonimizacao-no-enfrentamento-a-covid-19-e-o-privacywashing/>. Acesso em: 13 nov 2021.

RENZE, Matthew. Categorical vs. Numerical Data. Disponível em: <https://matthewrenze.com/articles/categorical-vs-numerical-data/>. Acesso em: 13 nov 2021.