FACULDADE DE TECNOLOGIA DE AMERICANA

Curso Superior de Tecnologia em Segurança da Informação

Rafael Figueiredo Caparroz

Anonimização e o projeto de Lei de Proteção de Dados Pessoais

Americana, SP

2016

FACULDADE DE TECNOLOGIA DE AMERICANA

Curso Superior de Tecnologia em Segurança da Informação

Rafael Figueiredo Caparroz

Anonimização e o projeto de Lei de Proteção de Dados Pessoais

Trabalho de Conclusão de Curso desenvolvido em cumprimento à exigência curricular do Curso Superior de Tecnologia em Segurança da Informação, sob a orientação do Prof. Benedito Aparecido Cruz

Área de concentração: Segurança da informação.

Americana, SP.

2016

FICHA CATALOGRÁFICA – Biblioteca Fatec Americana - CEETEPS
Dados Internacionais de Catalogação-na-fonte

C238q CAPARROZ, Rafael Figueiredo
Anonimização e o projeto de lei de proteção de dados pessoais. / Rafael Figueiredo Caparroz. – Americana: 2016.
53f.

Monografia (Curso de Tecnologia em Segurança da Informação). - - Faculdade de Tecnologia de Americana – Centro Estadual de Educação Tecnológica Paula Souza.
Orientador: Prof. Benedito Aparecido Cruz

1. Segurança em sistemas de informação I. CRUZ, Benedito Aparecido II. Centro Estadual de Educação Tecnológica Paula Souza – Faculdade de Tecnologia de Americana.

CDU: 681.518.5

Anonimização e o projeto de Lei de Proteção de Dados Pessoais

Trabalho de graduação apresentado como exigência parcial para obtenção do título de Tecnólogo em Segurança da Informação pelo CEETEPS/Faculdade de Tecnologia – FATEC/ Americana.
Área de concentração: Segurança da informação em bancos de dados

Americana, 06 de dezembro de 2016.

Banca Examinadora:

Benedito Aparecido Cruz (Presidente)
Professor
FATEC Americana

Edson Roberto Gaseta (Membro)
Professor
FATEC Americana

Rogério Nunes de Freitas (Membro)
Professor
FATEC Americana

AGRADECIMENTOS

Gostaria de agradecer aos meus pais, à minha irmã, e ao meu orientador, que me ajudaram nesta breve, mas intensa, jornada.

DEDICATÓRIA

Aos meus pais, que tanto me incentivaram durante a produção deste trabalho.

RESUMO

Atualmente, cada vez mais informação é coletada sobre indivíduos. A análise destes grandes volumes de informação possui muitas aplicações válidas. Entretanto, o potencial dano proporcionado pelo uso indevido de tais informações também é considerável. A anonimização dos dados, isto é, a remoção de informações pessoais de bancos de dados, é geralmente entendida como uma solução simples e elegante para este problema que supostamente preservaria a utilidade das informações e, ao mesmo tempo, a privacidade dos usuários, atingindo um equilíbrio ideal entre o interesse público e privado. Por esta razão, a anonimização de dados é uma prática disseminada. Entretanto, de acordo com vários estudos, procedimentos de anonimização não são tão eficazes como previamente aceito pela comunidade de segurança. O que estes estudos demonstram é que mesmo em bancos de dados supostamente expurgados de dados pessoais ainda é possível identificar parte dos usuários através da aplicação de certas técnicas de análise estatística. O presente trabalho busca descrever o debate acadêmico gerado por estes estudos e analisar como esta controvérsia sobre a anonimização de dados influenciou o debate realizado durante a consulta pública que resultou no projeto de lei de proteção de dados pessoais, atualmente em tramitação no congresso.

Palavras Chave: Direito à privacidade; Bancos de dados – medidas de segurança, Segurança em sistemas de informação.

ABSTRACT

Currently, more and more information is collected about individuals. The analysis of these large volumes of information have many valuable applications. However, the potential damage provided by the unauthorized use of such information is also considerable. The anonymization of data, i.e., the removal of personal information from databases, is generally understood as a simple and elegant solution to this problem that supposedly preserves the usefulness of the information and, at the same time, the privacy of users, reaching an ideal balance between public and private interest. For this reason, the anonymization of data is a widespread practice. However, according to several studies, anonymization procedures are not as effective as previously accepted by the security community. What these studies show is that even in databases supposedly expurgated of personal data it is still possible to identify the users by applying certain techniques of statistical analysis. This study aims to describe the academic debate generated by these studies and analyze how this controversy over the anonymization of data influenced the discussions during the public consultation which resulted in the personal data protection bill currently pending in Congress.

Keywords: *Right to privacy; Database security; Security in information systems.*

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIAÇÕES

ABRANET: Associação Brasileira de Internet

BRASSCOM: Associação Brasileira das Empresas de Tecnologia da Informação e Comunicação

GIC: *Group Insurance Commission*

GPoPAI-USP: Grupo de Pesquisa em Políticas Públicas para o Acesso à Informação da Universidade de São Paulo

HIPAA: *Health Insurance Portability and Accountability Act*

IMDB: *Internet Movie Database*

ITI: *Information Technology Industry Council*

MIT: *Massachusetts Institute of Technology*

PII: *Personally Identifiable Information*

SQL: *Structured Query Language*

1 INTRODUÇÃO

O funcionamento da sociedade moderna depende da coleta de volumes cada vez maiores de dados sobre seus membros. Graças ao poder da computação moderna, é possível utilizar este enorme volume de informações para uma variedade enorme de aplicações válidas: pesquisa de mercado, levantamentos sociológicos, desenvolvimento de tratamentos médicos, melhoria de políticas públicas, entre outras.

Entretanto, o potencial dano proporcionado pelo acesso a tais informações também é grande. Estas informações, quando relacionadas às pessoas a quem elas se referem, e acessadas por partes não autorizadas, fato que por si só já constitui uma violação da expectativa de privacidade do cidadão, podem ser utilizadas para fins de repressão, discriminação, retaliação e chantagem.

É por esta razão que uma lei de proteção de dados é necessária. Mais de cem países possuem legislação específica para a proteção de dados pessoais (GREENLEAF, 2015). O Brasil ainda não faz parte deste grupo, mas esta situação pode mudar em breve com a aprovação do projeto de Lei nº 5.276/2016, que atualmente tramita no congresso. A iniciativa regulatória foi elaborada através de uma consulta pública promovida pelo Ministério da Justiça iniciada em 28 de janeiro de 2015 e finalizada em 5 de julho do mesmo ano. O debate sobre a lei envolveu entidades do governo, da sociedade civil, do setor comercial e membros do público em geral. As contribuições foram coletadas no site criado no portal de participação do Ministério da Justiça. No dia 12 de maio de 2016, o texto gerado durante a consulta pública se transformou em projeto de lei.

Um dos aspectos que gerou mais controvérsia durante o debate realizada durante a consulta pública diz respeito a anonimização de dados. A anonimização de dados consiste na remoção de dados pessoais de um banco de dados. É um processo geralmente utilizado quando um administrador de dados deseja compartilhar um banco de dados com um terceiro que não deve ter acesso aos dados pessoais que estão contidos no banco.

A anonimização dos dados é apresentada como uma solução simples e elegante que une o melhor de dois mundos, permitindo que analistas possam encontrar informação útil e os usuários tenham sua privacidade preservada, atingindo um equilíbrio ideal entre o interesse público e privado. Por esta razão, a anonimização de dados é uma prática disseminada no âmbito da administração de bancos de dados.

Entretanto, de acordo com vários estudos, procedimentos de anonimização não são tão eficazes como previamente aceito pela comunidade de segurança. O que estes estudos demonstram é que mesmo em bancos de dados supostamente expurgados de dados pessoais ainda é possível reidentificar a maioria dos usuários através da aplicação de certas técnicas de análise estatística. Esta possibilidade existe porque a informação disponível no banco de dados anonimizado não é a única que existe sobre os indivíduos cujos dados estavam contidos no banco. Assim, é possível combinar a informação presente nos bancos “anonimizados” com fontes de dados externas para identificar novamente estes indivíduos. No campo de segurança da informação de banco de dados anonimizados, este ato é chamado de *reidentificação*.

O problema é agravado pela crescente disponibilidade de informações pessoais acessíveis pela Internet, parte delas publicadas pelos próprios indivíduos em redes sociais. Outra fonte possível de dados para um analista que busque reverter o processo de anonimização são outros bancos de dados anonimizados em que isto já aconteceu. Por esta razão, a reidentificação pode ser considerada um processo cumulativo.

Segundo Paul Ohm, professor de direito especializado em privacidade da informação, a anonimização de dados foi utilizada em substituição a uma análise cuidadosa dos riscos e benefícios envolvidos na divulgação de informações, pois supostamente eliminava os riscos à privacidade dos usuários. Esta distorção gerada pela crença na eficácia completa da anonimização teria produzido um desequilíbrio entre a preservação da privacidade e o interesse público. Nas palavras de Paul Ohm, "dados podem ser úteis ou perfeitamente anônimos, mas nunca podem ser ambas as coisas." (2010, p.1704). Esta controvérsia acadêmica sobre a eficácia da anonimização de dados influenciou bastante o debate sobre este tema durante a discussão do projeto de lei e será o foco deste trabalho.

O principal ponto de divergência entre os participantes da consulta pública residia na classificação legal dos dados anonimizados. Alguns dos participantes defenderam a tese de que, como o processo de anonimização pode ser revertido, dados anonimizados devem ser considerados dados pessoais e, portanto, devem estar sujeitos aos regulamentos estabelecidos pela Lei de Proteção de Dados Pessoais. Outros participantes discordaram desta posição, argumentando que dados

anonimizados não são dados pessoais e, consequentemente, não devem receber as proteções estabelecidas pela lei em discussão.

O objetivo deste trabalho é analisar como o debate acadêmico sobre a eficácia da anonimização influenciou a discussão deste tema durante a consulta pública sobre o projeto de lei de proteção de dados pessoais. A fim de atingir este objetivo, foram estabelecidos os seguintes objetivos específicos. O primeiro destes objetivos foi a realização de uma revisão bibliográfica sobre o assunto a fim de entender como é feita a anonimização, e como ela pode ser desfeita. O segundo objetivo foi a realização de uma pesquisa a respeito do estado atual da legislação brasileira sobre o tema da privacidade de dados pessoais. O último objetivo específico foi realização de um levantamento das posições dos participantes da consulta pública sobre o tema da anonimização.

A hipótese utilizada na produção deste trabalho é a de que os recentes desenvolvimentos nas técnicas de reidentificação levariam a uma reavaliação dos métodos de disseminação de informação, e que a controvérsia acadêmica gerada por esta reavaliação estaria presente no debate sobre o projeto de lei.

Neste trabalho, no primeiro capítulo será definido o que é e como é feita a anonimização dos dados. No segundo capítulo, será descrito como esta anonimização pode ser desfeita, usando como exemplos dois trabalhos de reidentificação de grande importância. Neste mesmo capítulo será discutido o debate acadêmico gerado por tais incidentes. No quarto capítulo, será delineado como esta controvérsia sobre a anonimização de dados influenciou o debate realizado durante a consulta pública que resultou no anteprojeto de lei de proteção de dados pessoais. Por fim, nas considerações finais deste trabalho, será realizada uma análise das partes relevantes do anteprojeto de lei apresentado, feita à luz dos mais recentes desenvolvimentos da ciência da reidentificação de dados.

Muitos dos termos técnicos utilizados neste trabalho são relativamente novos e, portanto, ainda não possuem correspondentes estabelecidos na língua portuguesa. Assim, sempre que possível, foi utilizada a tradução mais comum dos termos técnicos, seguidos, quando usados pela primeira vez no texto, pela expressão inglesa original entre parênteses. As definições destes termos técnicos geralmente acompanham a sua primeira utilização, mas também podem ser encontradas no glossário que acompanha este trabalho. Todos os termos presentes no glossário encontram-se em itálico quando utilizados pela primeira vez.

Outra observação relativa à terminologia utilizada neste trabalho se refere ao uso da palavra “anonimização” e seus derivados. O uso desta palavra é contestado por alguns especialistas no tema pois, segundo eles, ela leva o leitor a entender que o processo realmente torna anônimos os dados que passam pelo tratamento, algo que é contestado por estes especialistas. Paul Ohm, por exemplo, sugere o uso da palavra “*scrub*”, esfregar em inglês, pois esta indicaria apenas uma tentativa de limpar os dados pessoais de um banco de dados. Entretanto, dado que o termo anonimização é o mais utilizado na literatura científica, este também será utilizado neste trabalho, com a advertência que o termo deve apenas indicar a tentativa de tornar dados pessoais em dados anônimos, e não o seu sucesso completo.

Por fim, ainda sobre a terminologia utilizada, a expressão em inglês *Personally Identifiable Information* (PII) pode ser traduzida de várias formas. A expressão "*Informações Pessoais Identificáveis*” foi escolhida pois expressa melhor o conceito em português, além de ser mais comumente utilizada que as traduções alternativas em políticas de privacidade publicadas na internet, situação em que o conceito é geralmente utilizado. Em uma pesquisa realizada através da ferramenta de busca Google no dia 05 de novembro de 2016, uma consulta combinando a expressão “política de privacidade” com a expressão "Informações Pessoais Identificáveis" resultou em aproximadamente 29.600 resultados, mais do que a combinação de “política de privacidade” com as expressões “Informações de Identificação Pessoal” (28.700 resultados), Informações Pessoalmente Identificáveis” (11.900 resultados) ou "Informações Pessoais de Identificação" (8.180 resultados).

2 ANONIMIZAÇÃO: DEFINIÇÃO, APLICAÇÕES E MÉTODOS

Para que possamos entender como a anonimização pode ser desfeita, é necessário entender como ela é feita. Neste capítulo, será descrito o que é, para que serve e como é feita a anonimização.

2.1 Definição e aplicações

A anonimização de dados consiste no tratamento de bancos de dados para que, quando distribuídos a terceiros, já não contenham informação que permita o reconhecimento das pessoas a quem os dados se referem. Isto é feito através da remoção ou modificação dos dados pessoais presentes em um banco de dados. São considerados como dados pessoais quaisquer informações que permitam a identificação dos indivíduos no banco de dados, conhecidas no jargão da área como PII, sigla em inglês para *Personally Identifiable Information*, *Informações Pessoais Identificáveis*, em uma tradução livre.

Procedimentos de anonimização são amplamente utilizados porque supostamente permitem ao administrador de banco de dados preservar a privacidade dos indivíduos a quem estas informações se referem e, ao mesmo tempo, distribuir as informações contidas nestes bancos com o público, com outras organizações, ou mesmo com outros setores da mesma organização. Após ser anonimizada, a base de dados pode ser disseminada sem que haja a necessidade de onerosos controles sobre o uso destas informações.

Este modelo de compartilhamento de dados, também conhecido como "*publique e esqueça*", se baseia na suposição de eficácia do processo de anonimização destes dados, isto é, a hipótese de que um analista mal-intencionado, mesmo de posse de toda a base de dados distribuída, não seria capaz de reidentificar os indivíduos a quem eles se referem e utilizar esta informação de forma indevida.

A anonimização também é utilizada para fins de armazenamento de dados de longo prazo, de forma a evitar os custos relacionados ao controle de acesso que seriam incorridos caso a base ainda possuísse dados pessoais.

A anonimização de dados é uma prática amparada pela legislação de muitos países. A gestão de informações pessoais é normalmente sujeita a regulações bastante restritivas, entretanto, dados anonimizados são geralmente isentos de tais

restrições pela legislação. Um exemplo bastante influente de legislação que estabelece este tipo de exceção é a Diretiva Europeia Sobre Proteção de Dados Pessoais. A diretiva determina que cada país membro um deles edite leis sobre o processamento de “dados pessoais”, definidos como “qualquer informação relativa a uma pessoa singular identificada ou identificável” (UNIÃO EUROPEIA, Diretiva 95/46/EC, Art. 2º). Por extensão, dados devidamente anonimizados não se enquadram nesta definição e não são sujeitos aos mesmos regulamentos.

Outro exemplo importante são as cláusulas de privacidade constantes na lei federal americana Health Insurance Portability and Accountability Act (HIPAA), promulgada em 1996, que regulamenta os planos de saúde do país. Em 2003, a lei recebeu um adendo que estabelecia um padrão regulatório para o tratamento de dados médicos com o fim de proteger a privacidade dos pacientes.

As regras de privacidade da HIPAA estabelecem responsabilidades legais para administração de bancos de dados que contenham dados pessoais, mas isenta desta classificação toda informação que passa pelo um processo de anonimização que elimine ou modifique dezoito tipos de dados, listados no texto da legislação, que poderiam permitir a reidentificação de pacientes, desde que o administrador possua “conhecimento real que a informação restante pode ser usada, sozinha ou em combinação, para identificar o sujeito da informação” (ESTADOS UNIDOS DA AMÉRICA, 45 CFR 164.514, tradução minha).

A anonimização de dados é também uma prática que pode ser encontrada nas políticas internas de muitas empresas. A gigante de tecnologia Yahoo, por exemplo, dedica toda uma seção de sua política de privacidade aos aspectos legais do armazenamento e anonimização de dados. Nesta seção, a empresa garante aos usuários de seus serviços que os dados coletados sobre estes usuários serão anonimizados após 18 meses, contados a partir da data da coleta. A empresa também afirma que “toma medidas adicionais para que os dados coletados e usados para personalizar anúncios baseados em interesses (junto com algum conteúdo) no Yahoo não sejam associados a informações pessoais de identificação”, utilizando “um processo de várias etapas para substituir, truncar ou excluir identificadores para anular a identificação dos dados” (YAHOO! DO BRASIL INTERNET LTDA., 2016).

Empresas brasileiras também utilizam a anonimização. A seguradora Porto Seguro, por exemplo, afirma em sua política de privacidade que “poderá compartilhar os dados [coletados sobre os usuários] em referência com as demais empresas do

Grupo Porto Seguro e com terceiros de forma anonimizada", mas assegura o usuário que "emprega todos os esforços razoáveis de mercado para garantir a segurança de seus sistemas na guarda dos referidos dados" utilizando para isto os "métodos padrões e de mercado para anonimizar os dados coletados" (PORTO SEGURO SERVIÇOS E COMÉRCIO S.A., 2016).

A anonimização de dados é, portanto, uma solução de segurança da informação já consolidada, tanto pelo quadro jurídico quanto pela prática dos administradores de dados.

2.2 Métodos de anonimização

As principais técnicas de anonimização são as seguintes: supressão, substituição, generalização, perturbação e agregação. A fim de demonstrar a aplicação destas técnicas, será utilizado como exemplo uma base de dados fictícia, descrita pela tabela abaixo.

Tabela 1 – Exemplo de tabela não-anonimizada.

CPF	Nome	Data de Nascimento	Sexo	CEP	Renda Mensal	Data da Visita	Sintoma
185.302.491-00	José Augusto	25/05/65	M	13090-718	850,00	11/10/16	Falta de ar
271.132.131-20	Rafael Silva	12/06/65	M	13090-590	1200,00	12/10/16	Dor no peito
186.426.081-53	Maria Gomes	01/01/65	F	13080-379	3200,00	15/10/16	Pressão alta
098.705.201-20	Ataulfo Silva	02/06/65	F	13080-190	1900,00	18/10/16	Pressão alta
297.099.701-00	José Guerra	03/06/64	F	13080-340	900,00	21/10/16	Hemorragia
041.565.283-91	Mariana Góes	06/09/64	F	13080-350	1000,00	27/10/16	Fratura
258.842.631-04	Júlia Silva	03/09/64	M	13080-714	2500,00	02/11/16	Hemorragia
275.503.971-04	Rafael Ferreira	12/12/64	M	13080-360	1250,00	07/11/16	Pressão alta
098.705.201-20	Ataulfo Silva	02/06/65	F	13080-190	1900,00	07/11/16	Pressão alta
610.384.661-72	Mateus Ferreira	22/08/64	M	13080-691	1500,00	07/11/16	Dor no peito
296.416.591-20	Virgílio Alves	06/06/67	M	13080-369	900,00	09/11/16	Falta de ar
271.132.131-20	Rafael Silva	12/06/65	M	13090-590	1200,00	17/11/16	Pressão alta

Fonte: Tabela elaborada pelo autor.

2.2.1 Supressão

No exemplo acima, temos uma tabela contendo os registros de visitas ao setor de emergências do hospital. Evidentemente, dados como nome completo e CPF não

podem ser divulgados, pois permitem a identificação de um único indivíduo, sem necessidade de informação adicional. *Atributos* como estes são classificados como *Identificadores (Personal Identifiers)* e são considerados como um dos dois tipos de informações pessoais identificáveis. Também estão incluídos nesta classificação atributos que permitem uma comunicação direta com o indivíduo, como número de telefone e endereço de e-mail. Assim, tais dados precisam ser retirados da tabela, o que nos leva à primeira das técnicas de anonimização, a supressão.

A *supressão* consiste na remoção da informação pessoalmente identificável. No exemplo abaixo, os campos “Nome” e “CPF” são expurgados da tabela.

Tabela 2 – Exemplo de tabela após supressão.

Data de Nascimento	Sexo	CEP	Renda Mensal	Data da Visita	Sintoma
25/05/65	M	13090-718	850,00	11/10/16	Falta de ar
12/06/65	M	13090-590	1200,00	12/10/16	Dor no peito
01/01/65	F	13080-379	3200,00	15/10/16	Pressão alta
02/06/65	F	13080-190	1900,00	18/10/16	Pressão alta
03/06/64	F	13080-340	900,00	21/10/16	Hemorragia
06/09/64	F	13080-350	1000,00	27/10/16	Fratura
03/09/64	M	13080-714	2500,00	02/11/16	Hemorragia
12/12/64	M	13080-360	1250,00	07/11/16	Pressão alta
02/06/65	F	13080-190	1900,00	07/11/16	Pressão alta
22/08/64	M	13080-691	1500,00	07/11/16	Dor no peito
06/06/67	M	13080-369	900,00	09/11/16	Falta de ar
12/06/65	M	13090-590	1200,00	17/11/16	Pressão alta

Fonte: Tabela elaborada pelo autor.

Em alguns casos, a informação pessoal é criptografada, ao invés de ser excluída. Desta forma, o processo de anonimização pode ser revertido e os dados recuperados, mas somente por alguém que possua a chave utilizada.

2.2.2 Substituição

É também possível substituir os dados que possam levar à identificação de um indivíduo por outro tipo de informação, que não está diretamente relacionada ao indivíduo a quem se referem os dados. Em nosso exemplo, seria possível substituir o

CPF do paciente por um número identificador, o que preservaria a relação do paciente com suas visitas ao hospital, sem identificar a pessoa em questão. Isto seria útil caso um pesquisador quisesse, por exemplo, saber quantos pacientes retornam ao hospital em um certo período de tempo. Vejamos:

Tabela 3 – Exemplo de tabela após supressão e substituição.

Número de Usuário	Data de Nascimento	Sexo	CEP	Renda Mensal	Data da Visita	Sintoma
10001	25/05/65	M	13090-718	850,00	11/10/16	Falta de ar
10002	12/06/65	M	13090-590	1200,00	12/10/16	Dor no peito
10003	01/01/65	F	13080-379	3200,00	15/10/16	Pressão alta
10004	02/06/65	F	13080-190	1900,00	18/10/16	Pressão alta
10005	03/06/64	F	13080-340	900,00	21/10/16	Hemorragia
10006	06/09/64	F	13080-350	1000,00	27/10/16	Fratura
10007	03/09/64	M	13080-714	2500,00	02/11/16	Hemorragia
10008	12/12/64	M	13080-360	1250,00	07/11/16	Pressão alta
10004	02/06/65	F	13080-190	1900,00	07/11/16	Pressão alta
10009	22/08/64	M	13080-691	1500,00	07/11/16	Dor no peito
10010	06/06/67	M	13080-369	900,00	09/11/16	Falta de ar
10002	12/06/65	M	13090-590	1200,00	17/11/16	Pressão alta

Fonte: Tabela elaborada pelo autor.

No exemplo acima, é possível verificar que o paciente 10002 retornou ao pronto socorro do hospital pouco mais de um mês depois de sua primeira visita, com um sintoma diferente.

2.2.3 Generalização

Como será descrito no próximo capítulo, certos dados, como data de nascimento, sexo, raça e código postal, embora não identifiquem indivíduos de forma explicita, podem ser combinados de forma a identificar indivíduos em um banco de dados anonimizados. Isto porque, embora isoladamente um certo valor de um atributo possa ser comum à várias pessoas, em combinação com valores de outros atributos, muitas vezes esta combinação é única a um determinado indivíduo. Isto é um problema particularmente comum em bases de dados pequenas, como a do nosso exemplo. Dados assim são denominados *semi-identificadores*, e, juntamente com os

identificadores, compõem a totalidade do conjunto de informações pessoais identificáveis.

Um banco de dados devidamente anonimizados não deve conter dados pessoais, mas se todos os identificadores e semi-identificadores fossem removidos do banco de dados, geralmente os dados restantes não possuiriam muita utilidade. Este certamente seria o caso se isto ocorresse em nosso exemplo:

Tabela 4 – Exemplo de tabela com todas as Informações Pessoais Identificáveis suprimidas.

Sintoma
Falta de ar
Dor no peito
Pressão alta
Pressão alta
Hemorragia
Fratura
Hemorragia
Pressão alta
Pressão alta
Dor no peito
Falta de ar
Pressão alta

Fonte: Tabela elaborada pelo autor.

Uma forma de evitar situações como esta é através da generalização de dados. A *generalização* consiste na redução da precisão de certos dados, supostamente permitindo um equilíbrio entre utilidade e privacidade. No exemplo utilizado neste trabalho, se informações como data de nascimento e CEP fossem generalizadas, isto é, limitadas apenas ao ano de nascimento ou aos primeiros dígitos do código postal, estas seriam comuns a vários indivíduos, protegendo suas identidades e ao mesmo tempo permitindo o uso de dados demográficos e geográficos por analistas. Dados muito específicos normalmente não são necessários em análises estatísticas, que buscam identificar tendências gerais. Um exemplo de generalização pode ser encontrado na tabela seguinte.

Tabela 5 – Exemplo de tabela após supressão, substituição e generalização.

Número de Usuário	Ano de Nascimento	Sexo	CEP	Renda Mensal	Mês da Visita	Sintoma
10001	1965	M	13090	850,00	10/16	Falta de ar
10002	1965	M	13090	1200,00	10/16	Dor no peito
10003	1965	F	13080	3200,00	10/16	Pressão alta
10004	1965	F	13080	1900,00	10/16	Pressão alta
10005	1964	F	13080	900,00	10/16	Hemorragia
10006	1964	F	13080	1000,00	10/16	Fratura
10007	1964	M	13080	2500,00	11/16	Hemorragia
10008	1964	M	13080	1250,00	11/16	Pressão alta
10004	1965	F	13080	1900,00	11/16	Pressão alta
10009	1964	M	13080	1500,00	11/16	Dor no peito
10010	1967	M	13080	900,00	11/16	Falta de ar
10002	1965	M	13090	1200,00	11/16	Pressão alta

Fonte: Tabela elaborada pelo autor.

2.2.4 Perturbação

Outra forma de deixar os dados menos específicos e ao mesmo tempo preservar a utilidade destes dados é através da perturbação. Ao contrário do que ocorre quando os dados são generalizados, dados perturbados não sofrem apenas uma redução na precisão, eles são efetivamente alterados. Apesar disto, se realizada corretamente, a perturbação não altera de forma estatisticamente significante o conjunto dos dados, preservando assim a utilidade dos mesmos. A perturbação pode ser aplicada sobre a totalidade ou apenas parte dos dados. Uma das técnicas mais utilizadas de perturbação é a micro-agregação, que consiste na agregação de valores semelhantes e na substituição destes pela média dos valores.

Em nosso exemplo, o atributo “Renda Mensal” é o que melhor se adequa ao método de micro-agregação, pois é puramente numérico. Esta informação é bastante sensível, e não pode ser divulgada de maneira que possa identificar um indivíduo.

Inicialmente, é necessário remover registros adicionais de um mesmo usuário da tabela, pois presumivelmente possuem a mesma renda familiar. A tabela é então ordenada de acordo com este atributo, conforme ilustrado na tabela seguinte:

Tabela 6 – Exemplo de tabela ordenada de acordo com a renda mensal dos pacientes.

Número de Usuário	Ano de Nascimento	Sexo	CEP	Renda Mensal	Mês da Visita	Sintoma
10001	1965	M	13090	850,00	10/16	Falta de ar
10005	1964	F	13080	900,00	10/16	Hemorragia
10010	1967	M	13080	900,00	11/16	Falta de ar
10006	1964	F	13080	1000,00	10/16	Fratura
10002	1965	M	13090	1200,00	10/16	Dor no peito
10008	1964	M	13080	1250,00	11/16	Pressão alta
10009	1964	M	13080	1500,00	11/16	Dor no peito
10004	1965	F	13080	1900,00	10/16	Pressão alta
10007	1964	M	13080	2500,00	11/16	Hemorragia
10003	1965	F	13080	3200,00	10/16	Pressão alta

Fonte: Tabela elaborada pelo autor.

É feita então uma média dos valores de renda mensal de cada par de registros assim ordenados, conforme pode ser verificado na tabela abaixo:

Tabela 7 – Exemplo de tabela após a micro-agregação da renda mensal dos pacientes.

Número de Usuário	Ano de Nascimento	Sexo	CEP	Renda Mensal	Mês da Visita	Sintoma
10001	1965	M	13090	875,00	10/16	Falta de ar
10005	1964	F	13080	875,00	10/16	Hemorragia
10010	1967	M	13080	950,00	11/16	Falta de ar
10006	1964	F	13080	950,00	10/16	Fratura
10002	1965	M	13090	1225,00	10/16	Dor no peito
10008	1964	M	13080	1225,00	11/16	Pressão alta
10009	1964	M	13080	1700,00	11/16	Dor no peito
10004	1965	F	13080	1700,00	10/16	Pressão alta
10007	1964	M	13080	2850,00	11/16	Hemorragia
10003	1965	F	13080	2850,00	10/16	Pressão alta

Fonte: Tabela elaborada pelo autor.

Os registros são colocados novamente na ordem original, e os registros adicionais dos usuários são reinseridos na tabela, com os valores de renda mensal determinados pela média realizada acima. O resultado pode ser verificado na próxima tabela:

Tabela 8 – Exemplo de tabela após supressão, substituição, generalização e perturbação.

Número de Usuário	Ano de Nascimento	Sexo	CEP	Renda Mensal	Mês da Visita	Sintoma
10001	1965	M	13090	875,00	10/16	Falta de ar
10002	1965	M	13090	1225,00	10/16	Dor no peito
10003	1965	F	13080	2850,00	10/16	Pressão alta
10004	1965	F	13080	1700,00	10/16	Pressão alta
10005	1964	F	13080	875,00	10/16	Hemorragia
10006	1964	F	13080	950,00	10/16	Fratura
10007	1964	M	13080	2850,00	11/16	Hemorragia
10008	1964	M	13080	1225,00	11/16	Pressão alta
10004	1965	F	13080	1700,00	11/16	Pressão alta
10009	1964	M	13080	1700,00	11/16	Dor no peito
10010	1967	M	13080	950,00	11/16	Falta de ar
10002	1965	M	13090	1200,00	11/16	Pressão alta

Fonte: Tabela elaborada pelo autor.

2.2.5 Agregação

Neste modelo, a base de dados, mesmo “anonimizada”, não é distribuída a terceiros. Neste caso, informações são entregues aos analistas de forma agregada, em resposta a solicitações válidas. Por exemplo, caso um pesquisador precisasse saber quantas pessoas de mais de trinta anos foram atendidas no hospital com quadro de hemorragia, somente estas informações lhe seriam fornecidas pelo administrador do banco de dados.

Tabela 9 – Exemplo de resultado de uma consulta em uma base de dados agregada.

Consulta	Resultado
Número de pessoas de mais de 30 anos atendidas com quadro de hemorragia	2

Fonte: Tabela elaborada pelo autor

É importante notar que esta última técnica não se enquadra no modelo “publique e esqueça”, pois ela exige que o administrador de dados mantenha a posse dos dados completos e exerça controles sobre as consultas que lhe são dirigidas.

Neste capítulo, foi descrito como é feita a anonimização de dados. No próximo capítulo, será descrito como ela pode ser desfeita, e quais são as consequências disto.

3 COMO A ANONIMIZAÇÃO PODE SER REVERTIDA

Há um volume crescente de estudos científicos que contestam a eficácia das técnicas de anonimização em preservar a privacidade dos usuários. Trabalhos importantes demonstraram a possibilidade de reidentificar indivíduos cujos dados estavam contidos em bancos disponíveis ao público, através de informações que não foram expurgadas durante o processo de anonimização, por não serem consideradas Informações Pessoais Identificáveis. Dois experimentos de reidentificação possuem particular importância para este trabalho. O primeiro, realizado em 1997, levou ao desenvolvimento de um novo modelo de anonimização, que foi amplamente adotado; o segundo, realizado onze anos depois, colocou em questão a eficácia deste modelo.

3.1 A reidentificação dos dados do governador Weld

Em meados dos anos 90, no estado americano de Massachusetts, a agência responsável por administrar o plano de saúde dos funcionários públicos do estado, Group Insurance Comission (GIC), decidiu compartilhar os registros médicos destes funcionários com quaisquer pesquisadores que os requisitassem, sem custo. O então governador do estado, William Weld, assegurou que a privacidade dos funcionários seria preservada através da remoção dos dados pessoais do banco de dados a ser compartilhado.

Com o intuito de demonstrar a fragilidade destas práticas de anonimização, em 1997, Latanya Sweeney, na época estudante de pós-graduação no MIT, utilizou técnicas de análise estatística para identificar os registros médicos completos do governador do estado.

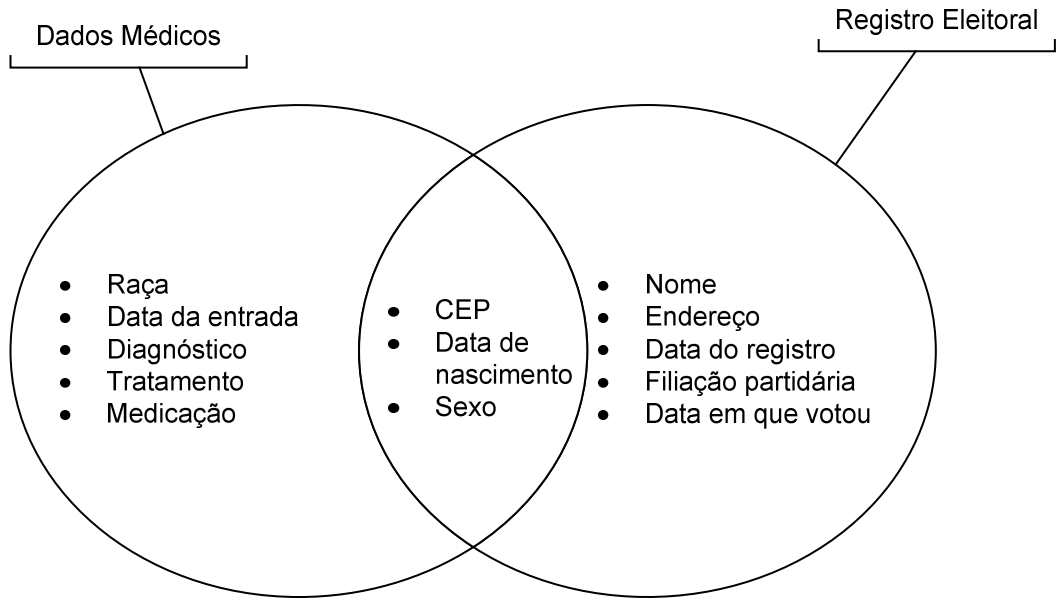
Um problema fundamental da prática de anonimização de dados é que as informações contidas na base de dados compartilhados não são as únicas disponíveis a um possível adversário, termo utilizado neste contexto para designar a pessoa, bem ou mal-intencionada, que busca identificar os indivíduos cujos dados estão contidos nesta base.

No caso da reidentificação dos dados do governador, estas informações foram obtidas através da combinação de informações ainda presentes no banco de dados compartilhado após o processo de anonimização com informações retiradas do

registro público de eleitores da cidade em que o governador residia, que Sweeney comprou legalmente por 20 dólares.

O registro eleitoral continha, entre outras informações, o nome completo de cada eleitor, juntamente com seu endereço, filiação partidária, data em que votou pela última vez, código postal, sexo e data de nascimento. Sweeney conseguiu reidentificar o governador porque estas últimas três informações - código postal, sexo, data de nascimento – também constavam, intactas, no banco de dados médico distribuído pelo GIC. Em linguagens SQL, esta operação de junção de duas tabelas mediante atributos comuns presentes em ambas é denominada *INNER JOIN*.

Figura 1: Diagrama de Venn representando a intersecção entre as duas fontes de dados.

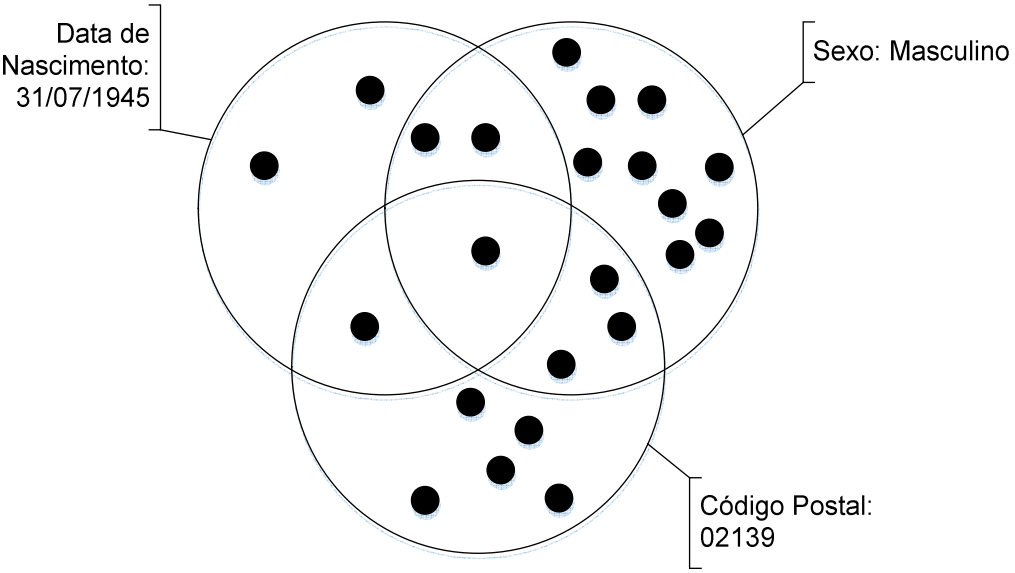


Fonte: Diagrama adaptado de SWEENEY, 2002, p 3.

Através da lista de eleitores, Sweeney conseguiu descobrir a data de nascimento e código postal do governador. Procurando no banco de dados distribuído pela GIC, a pesquisadora verificou que apenas um usuário do plano de saúde dos servidores possuía a mesma combinação de código postal, data de nascimento e sexo que o governador. A fim de confirmar a identidade do usuário, Sweeney analisou o outro conjunto de dados que possuía, a lista eleitoral da cidade de Cambridge. De acordo com esta lista, apenas seis pessoas na cidade dividiam a mesma data de aniversário, dessas pessoas, apenas três eram do sexo masculino, e, dentre esses homens, apenas um vivia na área coberta pelo seu código postal. Situações como esta não são incomuns: de acordo com a pesquisadora, 87% dos americanos podem

ser identificados através de apenas estas três informações (SWEENEY:2000, p. 17). Em um ato teatral, Sweeney entregou os registros médicos do governador no escritório da autoridade. Este processo de eliminação pode ser expresso pela tabela seguinte. Pontos representam indivíduos, mas as quantidades destes pontos, com exceção das mencionadas acima, são meramente ilustrativas.

Figura 2: Diagrama de Venn ilustrando o processo de reidentificação dos dados do governador William Weld.



Fonte: Diagrama elaborado pelo autor

Embora este caso tenha envolvido a reidentificação de uma única pessoa, as mesmas técnicas poderiam ter sido utilizadas para reidentificar grande parte dos usuários do banco de dados médicos distribuído pelo GIC, dada a frequência em que a combinação destes dados resultavam em um único registro.

As informações importantes na reidentificação dos dados do governador Weld, código postal e data de nascimento, não eram consideradas na época Informações Pessoais Identificáveis. Com base neste incidente, informações geográficas e de data foram incluídas na lista de informações que devem sofrer generalização segundo o HIPAA. De acordo com a lei, bancos de dados médicos anonimizados não devem conter subdivisões geográficas menores que um estado, o que equivale a apenas os três primeiros dígitos do código postal americano, e nenhuma informação de data mais especifica que a de ano. (ESTADOS UNIDOS DA AMÉRICA. 45 CFR 164.514)

3.2 Um novo modelo de anonimização: k-anonimidade

A fim de evitar a repetição de incidentes de reidentificação que utilizem as mesmas técnicas utilizadas no trabalho de reidentificação do governador Weld, Latanya Sweeney desenvolveu o conceito de *k-anonimidade* (k-anonimity) em um artigo publicado em 2002.

O objetivo do artigo de Sweeney é garantir cientificamente que, em um banco de dados anonimizado de acordo com o critério de k-anonimidade, não seja possível identificar os indivíduos a quem os dados se referem, e, ao mesmo tempo, preservar a utilidade dos dados.

O modelo de anonimização baseado na k-anonimidade é baseado em um pressuposto importante: o de que é possível ao administrador identificar corretamente semi-identificadores. Como foi visto no capítulo anterior semi-identificadores são atributos que, em combinação com outros atributos em uma fonte de dados externa, podem levar à reidentificação de um indivíduo. Consequentemente, se um administrador deixar identificar um atributo que também exista em uma fonte de dados externa, e que, portanto, possa servir de ligação entre os dois bancos, então a anonimização não será tão eficiente quanto deve ser. Sweeney admite que não é possível saber o que cada recipiente do banco e dados anonimizado possa saber, mas argumenta que “políticas e contratos, que se encontram fora dos algoritmos, podem ajudar” (SWEENEY, 2002, p. 8, tradução minha). Como será descrito nas seções seguintes, este pressuposto é contestado por alguns pesquisadores.

Segundo Sweeney, um banco de dados atende ao critério de k-anonimidade se, e apenas se, cada combinação de valores semi-identificadores que ocorra no conjunto de dados ocorrer no mínimo um número “k” de vezes. Consequentemente, o valor mínimo para “k” em que a privacidade é preservada é k=2, pois assim para cada combinação de valores semi-identificadores existem ao menos dois registros indistinguíveis um do outro. Esta condição é atingida através da supressão e generalização de dados identificadores e semi-identificadores. Um exemplo de tabela que atende ao critério de k-anonimidade é a tabela reproduzida na página seguinte.

A tabela 10 atende ao critério mínimo de k-anonimidade, com um índice de anonimidade k=2, considerando como semi-identificadores o conjunto de atributos

SI = {Raça, Ano de Nascimento, Sexo, CEP}. Quanto maior é o valor de k, mais difícil fica uma possível tentativa de reidentificação.

Tabela 10 – Exemplo de tabela que atende ao critério de k-anonimidade

Raça	Ano de Nascimento	Sexo	CEP	Sintoma
Negra	1965	M	13090	Falta de ar
Negra	1965	M	13090	Dor no peito
Negra	1965	F	13080	Pressão alta
Negra	1965	F	13080	Pressão alta
Negra	1964	F	13080	Hemorragia
Negra	1964	F	13080	Fratura
Branca	1964	M	13080	Hemorragia
Branca	1964	M	13080	Pressão alta
Branca	1964	M	13080	Dor no peito
Branca	1967	M	13080	Falta de ar
Branca	1967	M	13080	Fratura

Fonte: Tabela adaptada de SWEENEY, 2002.

O modelo de k-anonimidade, e seus derivados, foram amplamente adotados e se tornaram um padrão na indústria. Existem vários softwares de análise de anonimidade disponíveis no mercado baseados neste modelo. O próximo experimento, entretanto, levanta dúvidas sobre a eficácia deste modelo.

3.3 A reidentificação dos dados do prêmio Netflix

Em outubro de 2006, a Netflix, uma provedora *online* de filmes e séries de TV por assinatura, anunciou uma competição com o objetivo de desenvolver um melhor algoritmo de recomendação de filmes. A empresa ofereceu um prêmio de um milhão de dólares americanos para a primeira equipe que desenvolvesse um algoritmo mais eficiente que o então usado pela empresa, chamado *Cinematch*.

O algoritmo Cinematch funcionava analisando as avaliações de filmes de cada usuário para determinar uma lista de filmes similares aos que o usuário avaliou positivamente.

A fim de auxiliar os competidores no desenvolvimento do algoritmo, a empresa publicou um conjunto de dados anonimizado, contendo mais de cem milhões de

avaliações, feitas entre outubro de 1998 e dezembro de 2005, por mais de 480 mil usuários aleatoriamente selecionados. A única condição para a seleção dos usuários era que estes tivessem realizado mais de 20 avaliações de filmes. As avaliações mais recentes destes usuários foram omitidas.

De acordo com a empresa, para proteger a privacidade dos usuários do serviço, “toda informação pessoal identificando clientes individuais foi removida e identificadores de clientes [“*customer ids*”] foram substituídos por identificadores atribuídos de forma aleatória” (NETFLIX INC., 2006, tradução minha).

O conjunto de dados publicado após este processo de anonimização continha as seguintes informações: ID de usuário; nome do filme; data da avaliação e quantidade de estrelas dadas pelo cliente ao filme, um número inteiro de um a cinco. Além disso, a empresa perturbou parte dos dados publicados. Isto foi feito através da remoção, modificação e inserção de algumas das avaliações dos usuários. Apesar disto, segundo os responsáveis pela competição, o conjunto de dados publicado não era estatisticamente diferente dos dados originais.

De acordo com as regras do desafio, os competidores utilizariam os dados publicados para desenvolver um algoritmo que conseguisse prever como estes mesmos usuários classificaram os filmes nas avaliações mais recentes, que haviam permanecido secretas. A primeira equipe que desenvolvesse um algoritmo capaz de uma taxa de acerto 10% maior que o algoritmo *Cinematch* ganharia o prêmio.

Entretanto, dois pesquisadores, Armind Narayanan e Vitaly Shmatikov, enxergaram no prêmio Netflix uma oportunidade de provar a hipótese de que mesmo em conjuntos de dados como este, devidamente anonimizados e contendo apenas dados de comportamento, ainda é possível reidentificar usuários. Após meses de trabalho, a dupla publicou em 2008 um artigo comprovando esta hipótese. (NARAYANAN, A; SHMATIKOV, V, 2008).

Narayanan e Shmatikov descobriram que perfis de avaliação de filmes são surpreendentemente únicos e que bastam poucas informações externas para que se consiga identificar com sucesso com base apenas neste tipo de informação.

Isto acontece, segundo os autores, pois o conjunto de dados publicado pelo Netflix é um exemplo de *conjunto de dados de alta dimensão” (“high dimensional dataset”)*. Conjuntos deste tipo possuem um grande número de informações sobre cada indivíduo, de modo que o registro de cada indivíduo provavelmente é único. De fato, os dados publicados pelo Netflix se enquadram nesta classificação: de acordo

com os autores, o usuário médio nos dados publicados estava associado à 213 avaliações de filmes e datas em que estas foram realizadas.

Existem muitas formas de obter esta informação externa como, por exemplo, uma conversa aparentemente inocente sobre a opinião de uma pessoa sobre os filmes que ela assistiu recentemente. No caso específico do artigo, a dupla obteve esta informação externa através da análise de perfis públicos de usuários no site IMDB (*www.imdb.com*). O IMDB é um dos sites mais populares da Internet, uma espécie de enciclopédia virtual dedicada ao cinema e à TV com informações sobre quase 4 milhões de títulos. (IMDB, 2016). O site permite que usuários registrados publiquem resenhas e avaliações de filmes e programas de TV.

Porque os termos de serviço do IMDB proíbem a coleta de dados de forma automática, os pesquisadores coletaram manualmente informações sobre cinquenta usuários escolhidos aleatoriamente. Evidentemente, um adversário mal-intencionado poderia ter ignorado tal restrição, como notam os autores. Utilizando um algoritmo próprio, os autores conseguiram identificar dois dos cinquenta usuários do IMDB no conjunto de dados do Netflix. Por razões éticas, os pesquisadores não poderiam contatar os dois indivíduos para confirmar se as avaliações do Netflix foram realmente realizadas pelas mesmas pessoas, mas uma análise estatística determinou que a possibilidade de um falso positivo seria extremamente improvável.

Os autores também realizaram uma análise da eficácia do algoritmo que desenvolveram. Eles determinaram que sabendo apenas oito avaliações de filmes feitas por uma pessoa e data aproximada em que ela assistiu aos filmes (com uma margem de erro de 14 dias), mesmo com a possibilidade de que duas destas avaliações estejam completamente erradas, seria possível reidentificar 99% dos registros no banco de dados do Netflix. Até mesmo sabendo apenas duas avaliações, com datas com uma margem de erro de três dias, seria possível reidentificar 68% dos usuários.

Embora pareçam inócuos, dados como avaliações de filmes podem revelar informações que um indivíduo possa querer manter privadas. A forma como uma pessoa avalia certos filmes pode, por exemplo, revelar preferências sexuais e políticas da mesma. De qualquer forma, como notam os autores, mesmo que a maioria dos usuários do Netflix não se incomode em ter esse tipo de informação revelada, nenhum deles concordou expressamente em tornar estes dados públicos.

A mesma dupla realizou vários outros experimentos, demonstrando a possibilidade de reidentificação através da análise de relações de contatos em redes sociais, dados de localização agregados coletados por smartphones e hábitos de compra.

3.4 O fim do conceito de informação pessoal identificável?

Examinando as consequências dos seus experimentos de reidentificação, a mesma dupla de pesquisadores publicou em 2010 uma crítica do conceito de Informação Pessoal Identificável na revista “Communications of the ACM”, uma publicação importante no ramo da computação. No artigo, Narayanan e Shmaticov argumentam que os métodos de anonimização atuais, como k-anonimidade, encontram-se cada vez mais obsoletos. Estes métodos, como vimos, buscam inibir a reidentificação de bancos de dados através da modificação de atributos identificáveis, definidos previamente, de forma que atendam a certas condições que em tese impossibilitariam a combinação destes atributos com dados externos. O problema, de acordo com os autores, é que a anonimização de um conjunto pré-definido de atributos não é suficiente para impedir a combinação destes atributos com atributos de outros conjuntos de dados.

A causa fundamental deste problema estaria na distinção entre atributos identificáveis e não identificáveis, na qual se baseiam tais métodos de anonimização, distinção que, argumentam os autores, é arbitrária, e que faz “cada vez menos sentido à medida que a quantidade de informação disponível publicamente sobre indivíduos aumenta exponencialmente” (NARAYANAN, A; SHMATIKOV, V, 2010, p. 25, tradução minha).

Isto porque, ainda de acordo com os autores, qualquer informação que distingue uma pessoa de outra pessoa pode ser usada para fins de reidentificação. Assim, é possível identificar uma pessoa não apenas pelo que é, pelo tipo de atributo que geralmente é considerado como informação pessoal, mas também pelo que ela faz. Informações como históricos de pesquisa, relações de contatos, ou avaliações de filmes. Um analista que busque reidentificar um banco de dados anonimizados pode utilizar para este fim qualquer característica sobre uma pessoa que seja razoavelmente estável, e suficientemente variável de modo que seja improvável que duas pessoas possuam a mesma combinação de características.

Certamente, poucos considerariam dados como avaliações de filmes como informações pessoais, o que coloca o conceito em cheque. A cada estudo que comprova a vulnerabilidade de bancos anonimizados a este tipo de análise, encontrando novos tipos de dados passíveis de serem utilizados para o fim de reidentificação, o volume de dados a ser expurgado, modificado ou generalizado aumenta e, em razão inversa, a utilidade dos dados restantes diminui. Em face desta proliferação de dados potencialmente identificáveis, pesquisadores só poderão ter acesso a bases de dados cada vez mais editadas, e de menor valor científico.

Narayanan e Shmaticov reconhecem no artigo que tentativas de reidentificação baseadas na análise de atributos de comportamento exigem maior tolerância a erros e capacidade de computação e que, portanto, são mais caras e difíceis de implementar que análises baseadas apenas em informações pessoais. Entretanto, isto não é suficiente, segundo os autores, pois o avanço da tecnologia da computação e das técnicas de reidentificação facilitarão esforços de reidentificação baseados nesta abordagem.

Outro pesquisador com argumentos semelhantes é Paul Ohm, que em 2010 publicou uma análise das consequências legais dos recentes experimentos em reidentificação. Além dos incidentes de reidentificação dos dados do governador e dos dados do prêmio Netflix, Ohm cita o incidente de reidentificação dos dados de pesquisa dos usuários da AOL, ocorrido em 2006.

Neste incidente, a AOL, buscando se aproximar da comunidade de pesquisa, publicou um enorme conjunto de dados anonimizados, contendo mais de 20 milhões de consultas de pesquisa realizadas por mais de 650 mil indivíduos no serviço de buscas da empresa em um período de três meses. O banco de dados havia sido anonimizado através da substituição dos IPs e nomes de usuário por números identificadores, de forma a permitir que pesquisadores pudessem comparar os termos de pesquisa de cada usuário sem que, supostamente, pudessem identificar os indivíduos que fizeram estas buscas.

Infelizmente, esta suposição se demonstrou inválida. Michael Barbaro e Tom Zeller, repórteres do jornal americano New York Times, demonstraram isto ao identificar com sucesso um dos usuários presentes no banco de dados. A pessoa em questão, Thelma Arnold, uma mulher de 62 anos residente na cidade Lilburn, no estado da Geórgia, foi identificada pois fizera várias consultas por pessoas de mesmo sobrenome, além de pesquisas que revelaram sua localização, tais como “jardineiros

em Lilburn, Ga." ("*landscapers in Lilburn, Ga*"). O incidente causou grande embaraço para a AOL e resultou na demissão dos responsáveis pela publicação dos dados.

Além de ecoar os argumentos de Narayanan e Shmatikov em defesa do abandono do conceito de informação pessoal identificável, Paul Ohm argumenta que a reidentificação é um processo acumulativo, retroalimentar: cada caso de reidentificação com sucesso facilita futuras tentativas de reidentificação. Isto é, quanto mais informação sobre uma pessoa é revelada através da reidentificação de bancos de dados anonimizados mais fácil se torna a reidentificação desta pessoa em outros bancos de dados anonimizados, pois há mais informação que pode ser ligada a ela. Assim, mesmo a reidentificação de informação considerada inofensiva sobre uma pessoa aumenta a possibilidade de que informação sensível sobre esta pessoa possa ser revelada através da reidentificação de outros bancos de dados.

Segundo Paul Ohm, a reidentificação dos dados do Netflix é um bom exemplo deste princípio. Não só a revelação de quais filmes um usuário assistiu, e como os avaliou, pode ser usada contra este usuário, conforme citado acima, ela também abre espaço para revelações ainda mais danosas. De acordo com Paul Ohm, muitas pessoas utilizam o mesmo nome de usuário em vários sites. De posse do nome de usuário utilizado no IMDB, um adversário mal-intencionado poderia descobrir o nome de usuário usado pelo mesmo indivíduo no Facebook, por exemplo, e utilizar as informações ali presentes para responder às perguntas de segurança para a recuperação de senha de e-mail ou contas de banco. (OHM, 2010, p. 1748)

Paul Ohm também alude a um problema particularmente grave no modelo "publique e esqueça": o processo de anonimização pode ser reversível, mas a publicação dos dados não é. Uma vez que uma informação é publicada na internet, é praticamente impossível removê-la.

Ainda assim, o problema da reidentificação não poderia ser resolvido apenas evitando a liberação de dados anonimizados ao público. De acordo com Paul Ohm, ainda que a reidentificação de dados liberados ao público seja mais problemática, devido à razão citada acima, os recentes trabalhos de reidentificação não colocam apenas o modelo "publique e esqueça" em questão. Sempre que há o compartilhamento de dados anonimizados entre duas partes, existe o risco de reidentificação.

Por exemplo, um administrador de dados pode distribuir os dados anonimizados apenas para indivíduos ou organizações em que confia. Entretanto, um

destes indivíduos ou organizações poderia trair a confiança do administrador, e reidentificar os dados sem informá-lo.

Este risco existiria até mesmo em modelos de anonimização em que os dados não são entregues para um terceiro. Neste modelo, há um controle ativo: analistas enviam consultas ao administrador de dados e este apenas responde se o resultado não quebrar a privacidade dos usuários. Embora menos sujeito a uma tentativa de reidentificação massiva, porque um ataque deste tipo provavelmente seria detectado, modelos baseados na técnica de agregação também são suscetíveis a uma abordagem mais focalizada.

Paul Ohm oferece a hipótese que um adversário saiba que seu alvo deu entrada em um determinado hospital em uma determinada data, e que queira saber outras informações sobre ele. Para o administrador de dados, esta consulta poderia ser indistinguível de uma consulta legítima de um pesquisador médico.

3.5 Críticas e respostas

Os artigos citados na seção anterior fazem alegações bastante sérias, que colocam em questão práticas consolidadas no tratamento de dados pessoais. Como esperado, os artigos receberam bastante atenção da comunidade de segurança da informação. As reações foram em geral positivas, mas também geraram algumas críticas.

Dentre estas, vale destacar a crítica feita pelo pesquisador Khaled El Emam, especialista no tratamento de dados médicos eletrônicos. El Emam publicou em seu blog profissional um post crítico sobre o artigo de Paul Ohm argumentando que sua tese central, isto é, que os atuais métodos de anonimização não são efetivos, não é comprovada pela evidência apresentada (EL EMAN, 2009). Segundo El Eman, os exemplos de reidentificação citados por Ohm ocorreram em bancos de dados que não haviam sido apropriadamente anonimizados.

El Eman afirma que o incidente de reidentificação dos dados médicos do governador de Massachusetts ocorreu antes da promulgação do HIPAA, e não atende aos requisitos de privacidade exigidos por esta lei. Igualmente, de acordo com El Eman, a publicação das consultas de pesquisa dos usuários da AOL é outro exemplo de anonimização em que o administrador não usou as técnicas correntes, mesmo na época, de análise de risco e remoção de dados pessoais identificáveis.

Paul Ohm respondeu à estas críticas em uma versão posterior do artigo (OHM, p. 1727) e na seção de comentários do post no blog de El Eman, na qual os dois travaram uma discussão respeitosa. Ohm questionou no blog porque El Eman não mencionou o experimento de reidentificação dos dados do prêmio Netflix. Afinal, a forma como foram anonimizados os dados certamente atenderiam aos requisitos então correntes.

El Eman respondeu à resposta de Ohm, afirmando que não citou o trabalho de reidentificação dos dados do Netflix pois ele diz respeito a publicação de um conjunto de dados de alta dimensão. El Eman admitiu que a anonimização de conjuntos de dados desta natureza é particularmente difícil, mas também afirmou que existem métodos para a para a anonimização de bancos de dados deste tipo, e que novos estão sendo desenvolvidos. Por esta razão, argumentou o especialista, seria possível que os dados do Netflix não tivessem sido anonimizados corretamente.

Em resposta a isto, Ohm argumentou que, em todos os incidentes de reidentificação citados no artigo, os dados foram publicados por grandes organizações (GIC, AOL, Netflix) que em tese possuiriam políticas de segurança maduras e departamentos de TI bem organizados. Isto indica que a devida anonimização dos dados pessoais é um processo bastante complexo. Considerando isto, seria razoável supor que tais erros de anonimização seriam muito mais frequentes em organizações menores, com quadros de funcionários com pouco treinamento na prática na anonimização.

Ohm também argumenta, desta vez no artigo, que, nos casos de reidentificação citados, na época em que estes dados foram publicados, considerava-se que estes não continham dados pessoais. Foi somente após a comprovação da possibilidade de reidentificação que estes dados passaram a ser considerados dados pessoais. Desta forma, é possível que um em um banco de dados anonimizado de acordo com as regras atuais contenha informação que seja vulnerável à uma técnica de reidentificação ainda não descoberta. Assim, o exemplo de anonimização bem-feita de hoje pode ser o exemplo de anonimização malfeita de amanhã.

3.6 Alternativas ao modelo atual

Paul Ohm afirma que a legislação atual está defasada em relação ao corrente estado das técnicas de reidentificação. Segundo ele, as leis atuais podem ser

classificadas de acordo com a abordagem utilizada para a definição do conceito de dados pessoais. Esta definição pode ser feita de forma explícita, através da criação de listas de informações que devem ser consideradas como informações pessoais, como a HIPAA, ou de forma geral, definindo como informação pessoal toda informação que possa ser ligada à uma pessoa, como a diretiva europeia. Se as tendências atuais persistirem, então será necessária uma reavaliação destas leis. Isto porque, com a descoberta de novas técnicas que permitam a reidentificação através de novos tipos de informação, tais leis acabariam definindo dados pessoais de forma insuficientemente restritiva, ou de forma que englobariam todo tipo de informação, respectivamente. Nenhuma das alternativas seria desejável: a primeira não protegeria de forma satisfatória a privacidade dos cidadãos e a segunda criaria obrigações legais pesadas demais para praticamente qualquer administrador de dados.

Por esta razão, Ohm sugere uma estratégia em dois níveis: o estabelecimento de um padrão mínimo legal, para aplicação geral, e de regulamentações mais restritivas, em contextos em que o risco causado pela reidentificação é mais sério, como em relação ao trato de dados médicos ou financeiros, por exemplo.

Segundo Ohm, os problemas causados pelos recentes avanços na ciência da reidentificação exigem o abandono de abordagens baseadas apenas na identificação de dados pessoais. Ele sugere a substituição do modelo atual, puramente técnico, por um modelo com mais nuances, que também analise o risco causado por um potencial incidente de reidentificação.

De acordo com o pesquisador, legisladores não devem esperar o surgimento de uma solução técnica para este problema. Ao que tudo indica, nenhuma técnica de anonimização futura seria completamente eficiente, como se supunham ser as atuais, e antigas, técnicas de anonimização.

Ele também argumenta que a proibição do ato de reidentificação para fins não acadêmicos, embora desejável, não é suficiente. Isto porque a detecção de tais atos pode ser bastante difícil, uma vez que adversários mal-intencionados raramente anunciam suas ações.

Segundo o autor, a crença na efetividade da anonimização distorceu a análise dos riscos e benefícios envolvidos no compartilhamento de dados. Ohm também afirma que a análise dos riscos não deve se limitar ao exame dos possíveis danos causados pela revelação de dados pessoais, mas também sobre o quão é provável que tal evento ocorra. Assim, ele sugere a aplicação de um teste que considere cinco

fatores de risco. Ele afirma que esta lista não é exaustiva, e que outros fatores podem ser analisados de acordo com o caso a ser analisado.

O primeiro fator a ser avaliado diz respeito às técnicas utilizadas na anonimização dos dados. Embora estas nunca serão complemente eficientes, uma anonimização feita corretamente pode dificultar bastante o trabalho de um adversário.

O segundo refere-se à forma de publicação dos dados. A distribuição de dados ao público apresenta mais risco, pois é irreversível, enquanto que distribuição de dados apenas à pessoas e organizações confiáveis é mais seguro. Modelos em que não há distribuição de dados, nos quais o administrador de dados mantém o controle dos dados e responde a consultas de terceiros, são ainda mais seguros.

O terceiro fator a ser considerado tem relação com a quantidade de dados que serão distribuídos. Bancos de dados maiores facilitam o trabalho de um possível adversário e apresentam maior dano, caso este tenha sucesso.

O quarto fator que deve ser examinado diz respeito ao grau de motivação que a reidentificação de um banco de dados oferece a um possível adversário. Por exemplo, a reidentificação de um banco de dados médico pode ser bastante atraente para um administrador inescrupuloso de um plano de saúde, que queira saber quais possíveis clientes apresentam maior risco às finanças do plano.

O quinto, e último, fator que deve ser analisado tange ao grau de confiança que pode ser atribuído aos recipientes do banco de dados anonimizado. Um legislador poderia considerar, por exemplo, pesquisadores acadêmicos como mais merecedores de confiança, e ao mesmo tempo, atribuir menos confiança a empresas de publicidade direcionada.

Com base nesta análise dos fatores de risco, legisladores poderiam comparar os riscos e benefícios envolvidos no compartilhamento de dados, e legislar de acordo.

Neste capítulo, foi descrito como recentes experimentos de reidentificação abalaram a confiança nos atuais métodos de anonimização. No próximo capitulo, será analisado como este debate acadêmico influenciou o debate sobre este tema durante a consulta pública que gerou o projeto de lei nº 5.276/2016.

4 ANONIMIZAÇÃO E O PROJETO DE LEI DE PROTEÇÃO DE DADOS PESSOAIS

Neste capítulo serão analisados os aspectos legais da anonimização. A primeira seção faz um breve panorama do quadro jurídico brasileiro sobre a questão da privacidade de dados pessoais e descreve como foi produzido o projeto de lei. Na segunda, será realizado um breve resumo da lei, com atenção especial ao tema da anonimização. Por fim, a terceira seção deste capítulo trata sobre a controvérsia acerca do tema da anonimização que ocorreu durante o debate da lei.

4.1 Contexto legal

Embora o Brasil ainda não possua uma lei específica para a proteção dos dados pessoais, a privacidade destes dados é garantida pela constituição e tratados internacionais, de forma genérica, e por várias leis importantes, de forma dispersa e incompleta.

A Constituição Federal de 1988 estabelece em seu artigo 5º, inciso X, que “são invioláveis a intimidade, a vida privada, a honra e a imagem das pessoas, assegurado o direito a indenização pelo dano material ou moral decorrente de sua violação”. Esta proteção legal se desdobra nos dois incisos seguintes: o inciso XI, que garante a inviolabilidade do lar, e o inciso XII, que garante a inviolabilidade das comunicações.

A privacidade é direito assegurado também por vários tratados de direitos humanos dos quais o Brasil é signatário. Dentre estes tratados, cabe destacar a Declaração Universal dos Direitos Humanos, que em seu artigo XII, dispõe que “Ninguém será sujeito a interferências na sua vida privada, na sua família, no seu lar ou na sua correspondência, nem a ataques à sua honra e reputação. Toda pessoa tem direito à proteção da lei contra tais interferências ou ataques ” (ORGANIZAÇÃO DAS NAÇOES UNIDAS, 1948/2008, p. 15). Garantias semelhantes encontram-se em tratados como o Pacto Internacional dos Direitos Civis e Políticos, no âmbito da ONU, e a Convenção Americana de Direitos Humanos, no âmbito da Organização dos Estados Americanos (OEA).

Segundo o especialista em direito Fábio Condeixa (2015), tratados como estes “gozam de ‘status normativo supralegal’, isto é, na hierarquia normativa pátria, esses tratados estão acima das leis, sujeitando-se, no plano interno, apenas à CF.”

Algumas leis brasileiras também regulamentam o trato de informações pessoais, ainda que de forma secundária. Dentre elas, vale destacar as seguintes leis.

A lei nº 12.965, de 23 de abril de 2014, mais conhecida como o Marco Civil da Internet, é talvez a legislação mais relevante sobre o tema da privacidade de dados pessoais. Seu artigo 7º estabelece uma proteção sistemática da privacidade dos dados pessoais dos usuários da internet. Vejamos:

> "Art. 7º - O acesso à internet é essencial ao exercício da cidadania, e ao usuário são assegurados os seguintes direitos:
> I - inviolabilidade da intimidade e da vida privada, sua proteção e indenização pelo dano material ou moral decorrente de sua violação;
> II - inviolabilidade e sigilo do fluxo de suas comunicações pela internet, salvo por ordem judicial, na forma da lei;
> III - inviolabilidade e sigilo de suas comunicações privadas armazenadas, salvo por ordem judicial;
> [...]
> VII - não fornecimento a terceiros de seus dados pessoais, inclusive registros de conexão, e de acesso a aplicações de internet, salvo mediante consentimento livre, expresso e informado ou nas hipóteses previstas em lei;
> VIII - informações claras e completas sobre coleta, uso, armazenamento, tratamento e proteção de seus dados pessoais, que somente poderão ser utilizados para finalidades que:
> a) justifiquem sua coleta;
> b) não sejam vedadas pela legislação; e
> c) estejam especificadas nos contratos de prestação de serviços ou em termos de uso de aplicações de internet;
> IX - consentimento expresso sobre coleta, uso, armazenamento e tratamento de dados pessoais, que deverá ocorrer de forma destacada das demais cláusulas contratuais;
> X - exclusão definitiva dos dados pessoais que tiver fornecido a determinada aplicação de internet, a seu requerimento, ao término da relação entre as partes, ressalvadas as hipóteses de guarda obrigatória de registros previstas nesta Lei;
> XI - publicidade e clareza de eventuais políticas de uso dos provedores de conexão à internet e de aplicações de internet;
> [...]"

Vale destacar que os incisos VII e IX deste artigo prefiguram um princípio importante presente no projeto de Lei de Proteção de Dados Pessoais: o da proibição da transmissão de dados pessoais à terceiros sem o consentimento livre e informado do usuário a quem estes dados se referem. Outro aspecto importante do Marco Civil da Internet é que este também assegura o sigilo dos registros de conexão, o que inclui o IP do usuário, proibindo seu fornecimento à terceiros, salvo sob autorização judicial.

Outro marco importante na legislação brasileira com relação ao direto de privacidade dos dados pessoais foi estabelecido pela Lei de Acesso à Informação, sancionada em 2011. Segundo o texto da lei, a política do estado brasileiro com relação a informação gerada pela administração pública deve ser pautada pela transparência. Entretanto, é estabelecida uma exceção com relação aos dados pessoais coletados pelo governo. De acordo com a lei, “o tratamento das informações pessoais deve ser feito de forma transparente e com respeito à intimidade, vida privada, honra e imagem das pessoas, bem como às liberdades e garantias individuais." (BRASIL, Lei Nº 12.527, 2011, Art. 31). Por esta razão, o acesso aos dados pessoais deve ficar restrito aos agentes públicos autorizados por um prazo de até cem anos, salvo sob autorização do indivíduo ou nos casos estipulados pela lei.

Apesar das proteções legais já existentes, é ainda muito importante a existência de uma lei que estabeleça proteções específicas para os dados pessoais. Em tempos em que governos e empresas coletam volumes crescentes de informações sobre os cidadãos brasileiros, em que tais informações adquirem enorme importância para o funcionamento destas organizações, e em que o potencial uso indevido destas informações pode acarretar em grande dano para os indivíduos a quem elas se referem, é necessária uma lei específica, detalhada, para reger este assunto.

Razões econômicas também justificam a criação de uma lei de proteção de dados pessoais. Como afirmado na introdução deste trabalho, mais de cem países já possuem leis específicas para este fim e muitos deles não permitem que dados pessoais de seus cidadãos sejam armazenados ou transmitidos para países que em que tais proteções não estejam firmadas em lei específica.

O projeto de Lei de Proteção de Dados Pessoais teve sua origem em 2010, quando o Ministério da Justiça deu início à uma consulta pública que gerou a primeira versão do texto do projeto de lei. Esta primeira versão do texto foi então colocada em discussão no site do portal de participação do Ministério da Justiça em uma segunda consulta pública realizada entre 28 de janeiro de 2015 e 05 de julho do mesmo ano. A consulta pública recebeu mais de duas mil contribuições feitas por empresas, ONGs, instituições de ensino e pesquisa e de cidadãos, de forma individual. Estas contribuições foram realizadas através de comentários postados sobre os itens do texto apresentado no site ou através do envio de arquivos .pdf. Em 20 de outubro de 2015, o texto resultante foi apresentado ao público e, na Câmara dos Deputados, no dia 13 de maio de 2016, o texto se tornou o Projeto de Lei nº 5.276/2016.

No mesmo dia, o projeto de lei passou a tramitar em regime de urgência constitucional, o que significa que o projeto possui um prazo para ser votado, senão tranca a pauta da Câmara dos Deputados (CÂMARA DOS DEPUTADOS, 2016).

Em 18 de julho de 2016, a PL 5.276/2016 foi apensada ao projeto de lei nº 4060/2012, de autoria do deputado Milton Monti (PR/SP). Quando dois ou mais projetos de lei em tramitação no congresso legislam sobre temas semelhantes, o(s) projeto(s) mais novo(s) são apensados ao mais antigo, e passam a tramitar em conjunto. Os projetos devem então receber um parecer em conjunto. Se mais de um projeto receber um parecer favorável, o projeto mais antigo é substituído por uma versão única, composta pelos textos aprovados.

A PL 4060/2012 também dispõe sobre o tratamento de dados pessoais, mas o faz de forma menos detalhada. Enquanto a PL 5.276/2016 possui 56 artigos, que tomam mais de 18 páginas, a PL 4060/2012 possui 25 artigos, dispostos em seis páginas (MONTI, 2012). A PL 4060/2012 também não estabelece de forma clara a distinção entre dados anônimos e dados pessoais, razão pela qual decidi me focar no projeto de lei mais recente.

Os projetos de lei serão avaliados em comissão especial, constituída no dia 25 de outubro de 2016 (CÂMARA DOS DEPUTADOS, 2012). Se a PL 5.276/2016 receber um parecer favorável, será encaminhada para votação no plenário. Se aprovada, passa a tramitar no Senado.

4.2 Visão geral do projeto de lei

Nesta seção, será realizado um breve resumo da lei, com atenção especial ao tema da anonimização. Por razões de espaço, muitos detalhes não poderão ser citados aqui. O texto completo da lei pode ser encontrado no site da Câmara de Deputados, conforme indicado nas referências bibliográficas deste trabalho. (PODER EXECUTIVO, 2016)

O primeiro capítulo do projeto de lei dispõe sobre a aplicação e escopo da lei. Os artigos 1 e 3 determinam que a lei tem como objetivo regular o tratamento de dados pessoais coletados em território nacional, ou sobre indivíduos residentes no país, por pessoas físicas ou jurídicas, ainda que estas últimas tenham sede em outro país. O artigo 4 define que a lei não se aplica sobre dados coletados para fins jornalísticos,

literários e acadêmicos, assim como dados coletados para fins exclusivos de segurança pública e defesa nacional, que serão regulamentados por lei específica.

Neste capítulo também são estabelecidos os fundamentos (art. 2) e princípios (art.6) que devem nortear o tratamento de dados pessoais, assim como as definições para os termos utilizados no texto da lei (art. 5). Sobre este último artigo, vale destacar seus incisos I, IV e XII que definem “dados pessoais”, “dados anonimizados” e “anonimização”, respectivamente. Estas definições foram objeto de bastante controvérsia, como será descrito adiante. Segundo o mesmo artigo, a pessoa a quem se referem os dados tratados é denominada “titular” dos dados. Esta definição também será utilizada neste trabalho.

O segundo capítulo do projeto de lei estabelece os requisitos para o tratamento dos dados pessoais. Estes requisitos são estabelecidos no artigo 7, sendo que o mais importante destes é o consentimento livre e informado por parte do titular dos dados. No mesmo artigo também são definidos os casos em que este consentimento não é requerido. Para os fins deste trabalho, o mais relevante destes casos é o presente no inciso IV, que permite o tratamento de dados pessoais “para a realização de pesquisa histórica, científica ou estatística, garantida, sempre que possível, a anonimização dos dados pessoais”. Outra exceção relevante em que o consentimento do titular é dispensável é quanto o tratamento de dados é necessário para o atendimento da necessidade do próprio titular. Nestes casos, só devem ser coletados os dados estritamente necessários, sendo efetuada a anonimização destes dados sempre que possível (art. 10).

Os requisitos para o tratamento de dados sensíveis (arts. 11 e 12) são mais restritivos. Dados sensíveis são dados cujo uso indevido podem causar um dano maior para o titular, como por exemplo, os referentes à vida sexual e opiniões políticas e religiosas do indivíduo.

O artigo 8 garante ao titular acesso a informações detalhadas sobre o tratamento de seus dados. É importante notar que segundo o inciso V deste artigo, devem estar inclusas informações sobre os sujeitos para os quais estes dados pessoais podem ser transmitidos. O artigo 9 define as condições para a obtenção do consentimento do titular e estabelece que o titular tem o direito de revogar este consentimento.

O artigo 13 é bastante relevante para este trabalho, pois lida especificamente com o tratamento de dados anonimizados. Ele define que dados anonimizados devem

ser considerados dados pessoais quando “o processo de anonimização ao que foram submetidos for revertido ou quando, com esforços razoáveis, puder ser revertido”. O primeiro parágrafo deste artigo determina que dados utilizados para a produção de um perfil comportamental de uma pessoa, ainda que não identificada, devem ser considerados dados pessoais. O segundo parágrafo determina que um órgão competente pode estabelecer padrões e técnicas para a anonimização. O terceiro e último parágrafo estabelece que o uso e compartilhamento de dados deve ser feito de forma transparente, e que o órgão competente pode requisitar ao responsável “relatório de impacto à privacidade referente aos riscos de reversão do processo de anonimização e demais aspectos de seu tratamento”.

O terceiro capítulo estabelece os direitos do titular. Estes direitos, enumerados nos artigos 17 e 18, são: confirmação da existência dos dados; anonimização ou eliminação de dados desnecessários ou excessivos; acesso, correção, portabilidade e eliminação dos seus dados e, por fim, a aplicação das normas de defesa do consumidor. As formas como o titular pode acessar seus dados são detalhadas no artigo 19 e o artigo 20 estabelece que o titular tem direito a revisão de decisões tomadas automaticamente com base no tratamento de seus dados pessoais, desde que estas lhe afetem pessoalmente. Neste capítulo também é definido que os dados pessoais referentes ao exercício de direitos do titular não podem ser utilizados em seu prejuízo (art. 21), e que a defesa dos direitos estabelecidos por esta lei pode ser feita coletivamente (art.22).

O capítulo seguinte estabelece as obrigações legais do poder público com relação aos dados pessoais dos cidadãos. Os órgãos do governo só poderão coletar e tratar dados dos cidadãos à bem do interesse público (art. 23), fornecendo informações claras aos cidadãos sobre a forma do tratamento destes dados em veículos de fácil acesso (art. 24). Neste capítulo também é determinado que entidades do poder público só poderão transmitir dados pessoais para entidades privadas desde que isto seja estritamente necessário para a execução de uma atividade pública (art. 27) e que neste caso, é necessário coletar o consentimento do titular, exceto quando este é dispensado nos termos da lei.

O capítulo cinco trata da transferência de dados pessoais para outros países. Via de regra, ela será permitida quando no país destinatário existirem proteções no mínimo equiparáveis as estabelecidas por esta lei (art.33).

O sexto capítulo define quais são as responsabilidades das partes envolvidas no tratamento de dados. De acordo com o projeto de lei, o responsável é a pessoa, física ou jurídica, que toma as decisões relativas ao tratamento dos dados; e o operador é a pessoa, física ou jurídica que efetivamente realiza este tratamento (art. 38). Também existe a figura do encarregado, que é uma pessoa física indicada pelo responsável para atuar como um canal de comunicação com os titulares e também com o poder público (art. 41).

O sétimo capítulo estabelece que o operador do tratamento de dados pessoais deve seguir regras de segurança adequadas de forma a prevenir incidentes de acesso não autorizado, perda, alteração e divulgação de dados pessoais (art. 45). Caso tal incidente ocorra, o responsável deve contatar prontamente o órgão competente (art.47), que determinará as providências a serem tomadas (art.48).

O oitavo capítulo determina quais são as possíveis punições aos quais estão sujeitos os responsáveis em caso de infração de alguma das regras estabelecidas pela lei (art. 52). Também é definido neste capítulo quais serão as atribuições do órgão administrativo a ser criado para garantir a correta aplicação da lei (art. 53).

Por fim, no último capítulo, é definido que a lei entrará em vigor no prazo de 180 dias a partir da data de sua publicação.

4.3 Controvérsias sobre a anonimização

O debate acadêmico sobre a eficácia das atuais técnicas de anonimização influenciou bastante o debate sobre o projeto de lei durante a consulta pública realizada em 2015.

A principal divergência entre os participantes dizia respeito à classificação legal dos dados anonimizados. Alguns dos participantes argumentaram que dados anonimizados não deveriam ser considerados dados pessoais, e que, portanto, se encontrariam fora do escopo da lei. Outros participantes, entretanto, argumentaram que dados anonimizados deveriam ser considerados dados pessoais pois o processo de anonimização pode ser revertido. A primeira posição foi geralmente defendida por empresas de telecomunicações e associações da indústria de tecnologia, como a Cisco, a ABRANET (Associação Brasileira de Internet), a Brasscom (Associação Brasileira das Empresas de Tecnologia da Informação e Comunicação) e a ITI (*Information Technology Industry Council*). Já a segunda posição foi defendida em

geral por integrantes de organizações de defesa do consumidor, organizações de defesa da privacidade e centros de pesquisa acadêmicos, tais como a associação PROTESTE, o coletivo Intervozes e o Grupo de Pesquisa em Políticas Públicas para o Acesso à Informação da Universidade de São Paulo (GPoPAI-USP).

Tabela 11: Tabela relacionando participantes da consulta pública de acordo com sua posição sobre a classificação dos dados pessoais.

Expansão do conceito de dados pessoais	Restrição do conceito de dados pessoais
Centro de Tecnologia e Sociedade da FGV - Direito/RIO	ABRANET (Associação Brasileira de Internet)
Bruno B. Bioni	BRASSCOM (Associação Brasileira das Empresas de Tecnologia da Informação e Comunicação)
GPoPAI-USP (Grupo de Pesquisa em Políticas Públicas para o Acesso à Informação da Universidade de São Paulo)	Cisco
Intervozes	ITI (Information Technology Industry Council)
PROTESTE	RELX Group
Coletivo Antivigilância	US Business Council

Fonte: Elaborada pelo autor

Em termos gerais, esta divergência se traduziu em disputas a respeito da redação de três incisos presentes no artigo 5 do projeto de lei, que estabelece as definições dos termos utilizados no texto da lei. O texto apresentado para debate continha as seguintes definições:

> "Art. 5º Para os fins desta Lei, considera-se:
> I - dado pessoal: dado relacionado à pessoa natural identificada ou identificável, inclusive a partir de números identificativos, dados locacionais ou identificadores eletrônicos;
> [...]
> IV – dados anônimos: dados relativos a um titular que não possa ser identificado, nem pelo responsável pelo tratamento nem por qualquer outra pessoa, tendo em conta o conjunto de meios suscetíveis de serem razoavelmente utilizados para identificar o referido titular;
> [...]
> XIV – dissociação: ato de modificar o dado pessoal de modo a que ele não possa ser associado, direta ou indiretamente, com um indivíduo identificado ou identificável;
> [...]" (MINISTÉRIO DA JUSTIÇA, 2015)

A definição de dados pessoais (e por extensão, a de dados anônimos) é particularmente importante pois ela define o escopo da lei. Durante a consulta pública, houve intenso embate entre aqueles que buscavam expandir esta definição, de forma a incluir explicitamente dados anonimizados, e aqueles que buscavam restringi-la, de forma a deixar claro que dados anônimos não deveriam ser considerados dados pessoais.

Praticamente todos os defensores da expansão da definição citaram os trabalhos de reidentificação discutidos no capítulo anterior de forma a corroborar esta tese. Por exemplo, na contribuição feita pelo Centro de Tecnologia e Sociedade da FGV- Direito/RIO, o trabalho de Latanya Sweeney foi utilizado como um exemplo de como “dados inicialmente não definidos como dados pessoais podem ser utilizados para identificar um indivíduo” (p. 10). Citando Paul Ohm, os pesquisadores da FGV argumentaram que “a ciência da reidentificação demonstrada em estudos recentes minou a fé depositada nas técnicas de dissociação, de forma que a partilha de dados de forma indiscriminada e o armazenamento perpétuo de dados já não se justifica pela garantia de privacidade” (p. 10). Por esta razão, os integrantes do centro de pesquisa sugerem que fique explícito no projeto de lei que dados anônimos são considerados dados pessoais. Para este fim, os pesquisadores da FGV adotaram a mesma sugestão de alteração promovida pelos pesquisadores da GPoPAI-USP.

Os pesquisadores da USP sugeriram a substituição do termo “dissociação” por “anonimização”, pois segundo eles a dissociação é apenas uma das técnicas utilizadas no processo de anonimização. Segundo a classificação utilizada no primeiro capítulo deste trabalho, a dissociação englobaria apenas as técnicas de supressão e substituição, mas não as demais. Joana Varon, integrante da oficina Antivigilância, defendeu a mesma alteração e também sugeriu a substituição do termo “dados anônimos” pelo termo “dados anonimizados”, a fim de indicar que o processo de anonimização pode ser revertido. Os pesquisadores da USP recomendaram a eliminação do inciso que definia dados anônimos, pois estes estariam inclusos no conceito de dados pessoais, e sugeriram a seguinte definição para o termo “anonimização”:

> “IV – anonimização: ato de tornar um dado não correlacionável ao seu titular, utilizando-se de técnicas que procurem não identificá-lo, direta ou indiretamente, com um indivíduo. Os dados anônimos são, para fins desta lei, dados pessoais em razão da possível reversibilidade de seu

> processo, ainda que disponha de regras próprias nos termos desta legislação; ” (GPoPAI-USP, 2015, p.11)

Os pesquisadores da USP reconheceram que, embora a anonimização possa ser revertida, ela representa uma prática de segurança que deveria ser encorajada. Assim, defenderam a inclusão dos dados anonimizados no conceito de dados pessoais, mas também defenderam que estes dados anonimizados deveriam estar sujeitos a regras mais flexíveis que as que regulam demais dados pessoais. Neste sentido, sugeriram que a exigência do consentimento do titular fosse dispensada para o tratamento de dados anonimizados, desde que certos requisitos fossem atendidos.

Estes requisitos foram considerados pelos pesquisadores como essenciais para que o regime legal mais flexível garantido pela anonimização dos dados não acabasse esvaziando o projeto de lei. O primeiro requisito é que o processo de anonimização seguisse padrões estabelecidos pelo órgão competente, que fiscalizaria sua correta aplicação. Outro requisito é a elaboração pelo responsável pela anonimização de relatórios de impacto de privacidade, em caso de compartilhamento de dados. Os pesquisadores também defenderam a obrigatoriedade de autorização pelo órgão competente em caso de distribuição de dados anonimizados ao público. Neste caso, o órgão competente avaliaria se os riscos incorridos pela publicação dos dados são aceitáveis ou não. Por fim, os pesquisadores da USP recomendaram a proibição da reidentificação de dados anonimizados, salvo consentimento do titular.

Bruno R. Bioni, pesquisador visitante no Centro de Tecnologia, Sociedade e Direito da Universidade de Ottawa, ecoou esta posição em um comentário sobre a definição de “dados anonimizados”, citando também o trabalho de Narayanan e Shmatikov e acrescentando que, em “tempos de Big Data e *data agregation* torna-se cada vez mais porosa a dicotomia conceitual de dados anônimos e dados pessoais e, em última análise, o que deve ou não ser considerado como uma prática razoável para "desanonimizar" um dado” (BIONI, 2015).

No outro lado da controvérsia, entidades ligadas ao setor empresarial defenderam uma definição mais restritiva de dados pessoais. Quase todos os participantes que defendiam esta posição recomendaram que o conceito de informação identificável presente na definição de dados pessoais estivesse sujeito ao critério de razoabilidade.

O Information Technology Industry Council (ITI), entidade que possui entre seus membros empresas como Google, Samsung, Facebook, HP e IBM, argumentou em

sua contribuição ao debate que a definição de dados pessoais no anteprojeto de lei era abrangente demais, pois, da maneira que estava redigida, incluiria até mesmo dados que não poderiam ser razoavelmente relacionados aos seus titulares. Por esta razão, defendia a exclusão explícita dos dados anonimizados desta classificação (ITI, 2015, p. 4-5). A Brasscom, Associação Brasileira das Empresas de Tecnologia da Informação e Comunicação, defendeu a mesma tese, argumentando que uma definição mais restritiva, como a presente na legislação canadense, seria mais adequada para equilibrar a proteção do titular com o livre fluxo de informações. A Brasscom também alegou que, em discussões recentes sobre o âmbito da aplicação da lei europeia, tem sido sugerida uma definição mais restritiva de dados pessoais (BRASSCOM, 2015, p. 7). Também defendendo a mesma tese, o RELX Group, grupo atuante no setor de captação, desenvolvimento e gerenciamento de informações, afirmou em sua contribuição que "há um número crescente de casos nos quais a análise de dados [anonimizados] leva a descobertas importantes que beneficiam a sociedade em geral" (RELX GROUP, 2015, p. 5). Esta empresa sugeriu a seguinte redação para o inciso I do artigo 5:

> "I – dado pessoal: dado relacionado à pessoa natural identificada ou razoavelmente identificável, inclusive a partir de números identificativos, dados locacionais ou identificadores eletrônicos;
> Parágrafo 1 – dados anônimos ou descaracterizados estão excluídos desta lei. " (RELX GROUP, 2015, p. 5)

A Brasscom também argumenta que, como a lei não se aplica ao tratamento de dados anônimos, este poderia "ser efetuado sem quaisquer exigências ou formalidades" (BRASSCOM, 2015, p. 9). Além disso, ela sugeriu que a definição de dados anônimos não fizesse menção à possibilidade de identificação destes dados por terceiros, pois "o responsável desconhece que outros meios podem ser empregados por terceiros ('outras pessoas') em tentativas de re-identificação de dados anônimos". Assim, bastaria que o responsável pelos dados não fosse capaz de identificar os dados para que estes fossem considerados dados anônimos. Para este fim, a Brasscom sugeriu a seguinte redação para o inciso IV do artigo 5 (palavras taxadas indicam sugestões de trechos a serem eliminados, palavras sublinhadas indicam sugestões de adições ao projeto de lei):

> “IV - dados anônimos: dados ~~relativos~~ ~~a~~ sobre um titular que não possa ser identificado, ~~nem~~ pelo responsável pelo tratamento ~~nem por qualquer outra pessoa~~, tendo em conta o conjunto de meios suscetíveis de serem razoavelmente utilizados para identificar o referido titular. ” (BRASSCOM, 2015, p. 9, grifos no original).

Os defensores da posição que dados anonimizados não devem ser considerados dados pessoais não citaram os trabalhos de reidentificação citados no capítulo anterior, mesmo que para refutá-los. Implicitamente, entretanto, a maioria destes participantes reconhecem a possibilidade de reidentificação, razão pela qual pressionaram pela inclusão da clausula de razoabilidade da potencial identificação de uma pessoa. Seguindo este argumento, tais participantes consideram a possibilidade de reidentificação de dados devidamente anonimizados uma possibilidade remota, não razoável.

O texto produzido no final da consulta pública não deixou explicito, nas definições constantes no art. 5, se os dados anônimos seriam considerados dados pessoais ou não. Vejamos:

> “Art. 5º Para os fins desta Lei, considera-se:
> I - dado pessoal: dado relacionado à pessoa natural identificada ou identificável, inclusive números identificativos, dados locacionais ou identificadores eletrônicos quando estes estiverem relacionados a uma pessoa;
> [...]
> IV - dados anonimizados: dados relativos a um titular que não possa ser identificado;
> [...]
> XII - anonimização: qualquer procedimento por meio do qual um dado deixa de poder ser associado, direta ou indiretamente, a um indivíduo;
> [...]”

Conforme pode ser visto acima, a definição de dados pessoais no artigo 5 ficou um pouco mais restritiva, mas não incorporou a cláusula de razoabilidade, e tampouco uma exclusão explícita dos dados anonimizados, como defendiam a Brasscom, a ITI e o RELX Group. Por outro lado, a sugestão de redação feita pelo grupo da USP também não foi aceita, e a definição de anonimização não incluiu o parágrafo que estipulava que dados anonimizados são dados pessoais. Entretanto, as duas mudanças na terminologia sugeridas por Joana Varon foram adotadas. A definições de dados anonimizados e anonimização ficaram mais curtas, menos específicas, mas não sofreram uma mudança significativa.

O texto final do projeto de lei ganhou um artigo específico sobre o tratamento de dados anonimizados, artigo que não existia no projeto original. O artigo em questão segue abaixo em sua íntegra.

> “Art 13. Os dados anonimizados serão considerados dados pessoais para os fins desta Lei quando o processo de anonimização ao qual foram submetidos for revertido ou quando, com esforços razoáveis, puder ser revertido.
> § 1º Poderão ser igualmente considerados como dados pessoais para os fins desta Lei os dados utilizados para a formação do perfil comportamental de uma determinada pessoa natural, ainda que não identificada.
> § 2º O órgão competente poderá dispor sobre padrões e técnicas utilizadas em processos de anonimização e realizar verificações acerca de sua segurança.
> § 3º O compartilhamento e o uso que se faz de dados anonimizados deve ser objeto de publicidade e de transparência, sem prejuízo do órgão competente poder solicitar ao responsável relatório de
> impacto à privacidade referente aos riscos de reversão do processo de anonimização e demais aspectos de seu tratamento.”

A leitura do artigo 13 indica que dados anonimizados não são considerados dados pessoais, e sugere que o critério para esta distinção reside na razoabilidade de reversão do processo de anonimização. Neste aspecto, pode-se entender que as recomendações feitas pelos defensores de um conceito mais restritivo de dados pessoais foram adotadas.

Entretanto, o processo de anonimização de dados foi regulamentado, contra os desejos da Brasscom. Alguns dos requisitos para a anonimização propostos pelos pesquisadores da USP foram adotados, embora de forma atenuadas. O órgão competente recebeu poderes de regulamentação e fiscalização e poderia, a seu critério, exigir relatórios de impacto à privacidade.

5 CONSIDERAÇÕES FINAIS

Como vimos no capítulo anterior, o artigo de Paul Ohm foi frequentemente citado por participantes da consulta pública que defendiam uma regulamentação mais ativa sobre o tratamento de dados anonimizados. Isto é de se esperar, pois se trata de um artigo bastante detalhado, mas ao mesmo tempo, escrito em linguagem acessível à leigos.

Entretanto, da forma em que o artigo está escrito, ele pode induzir o leitor a entender que a reidentificação é um processo mais fácil do que ele é na realidade. Isto ocorre porque falta ao artigo uma distinção mais clara entre os dois principais incidentes de reidentificação citados. Conforme notou corretamente Khaled El Eman, o caso da reidentificação dos dados do governador Weld não é um bom exemplo de vulnerabilidade a este tipo de ataque pois o tratamento dos dados médicos feito pela GIC não se conforma aos padrões atuais de anonimização. De fato, foi com base nesta experiência que Latanya Sweeney desenvolveu o conceito de k-Anonimidade, que, quando devidamente aplicado, é bastante efetivo na prevenção de ataques desta natureza.

Paul Ohm usa várias vezes em seu artigo a expressão “fácil reidentificação” e sobre este aspecto afirma o seguinte:

> “O estudo Netflix revela que é surpreendentemente fácil reidentificar pessoas em dados anonimizados. Embora o usuário médio de computador não saberia realizar um “*inner join*”, a maioria das pessoas que fizeram um curso de gerenciamento de banco de dados ou trabalhou em TI provavelmente poderia replicar essa pesquisa usando um computador rápido e softwares amplamente disponíveis como o Microsoft Excel ou Access. ” (OHM, 2010, p. 1730, tradução minha)

Entretanto, este argumento poderia ser contestado. A reidentificação dos dados do governador William Weld, poderia ser considerada relativamente fácil, acessível a qualquer analista que saiba fazer um *“INNER JOIN”*, mas o trabalho envolvido na reidentificação dos dados do prêmio Netflix é algo que exige considerável conhecimento e esforço. Conforme afirmado pelos próprios pesquisadores responsáveis por este trabalho, “algoritmos de reidentificação baseados em atributos comportamentais”, como o usado na reidentificação dos dados do Netflix, “devem tolerar uma certa imprecisão nos valores dos atributos. Eles são, portanto, computacionalmente mais caros e mais difíceis de implementar do que a

reidentificação com base em semi-identificadores demográficos". (NARAYANAN, A; SHMATIKOV, V, 2010, p. 26, tradução minha)

Apesar disto, esta objeção não invalida a tese central de Paul Ohm, a de que os recentes avanços das técnicas de reidentificação devem levar a uma reavaliação das atuais práticas de compartilhamento de dados pessoais. Isto porque, embora a reidentificação de bancos de dados através da análise de dados comportamentais seja mais difícil de ser feita, ela é uma realidade e apresenta um risco considerável para a privacidade dos usuários. Mesmo defensores das atuais técnicas de anonimização, como El Eman, admitem que elas não são muito adequadas para dados desta natureza. Além disso, conforme notam Narayanan e Shmatikov, fatores como avanços nas técnicas de reidentificação, incentivos financeiros cada vez maiores para potenciais adversários e a quantidade crescente de informação disponível na internet sobre as pessoas tendem a facilitar no futuro tentativas de reidentificação através deste tipo de abordagem.

Também deve causar preocupação o fato que incidentes de reidentificação ocorreram mesmo com dados publicados por grandes organizações, que em tese possuiriam políticas de segurança bem desenvolvidas e departamentos de TI bem operados. Isto indica que a devida anonimização dos dados pessoais é um processo igualmente complexo. Se a anonimização não é feita corretamente, pode-se sim argumentar que é fácil a reidentificação destes dados. Por esta razão, é positiva a determinação, constante no texto do projeto de lei, de que o órgão competente tem a capacidade de estabelecer padrões para a anonimização de dados, e de fiscalizar sua correta aplicação.

Por tais razões, é bastante válida a posição de Paul Ohm que é necessário que o atual modelo de anonimização, baseado apenas na análise técnica dos dados, seja substituído por um modelo mais completo, que envolva uma análise dos riscos e benefícios envolvidos no compartilhamento de dados. Assim, pode-se afirmar que as recomendações feitas pelos participantes que defendiam que dados anonimizados deveriam estar dentro do escopo da lei eram as mais adequadas, tendo em vista a literatura científica atual.

Como foi descrito no capítulo anterior, esta definição ampliada de dados pessoais não foi incorporada na versão final da lei. Ainda assim, muitas das recomendações destes participantes sobre os critérios para a anonimização foram adotadas no artigo 13 do projeto de lei. É possível afirmar que as normas presentes

neste artigo estabelecem proteções suficientes contra o risco de reidentificação, se implementadas corretamente. A efetiva aplicação do das normas presentes projeto de lei depende principalmente da atuação do órgão competente a ser criado. Assim, caso o projeto de lei seja aprovado, o órgão competente deve fazer jus a seu nome.

REFERÊNCIAS BIBLIOGRÁFICAS

BIONI, B. **Contribuição em comentário sobre o inciso IV do artigo 5 do texto do Anteprojeto de Lei de Proteção de Dados Pessoais**. 2015. Disponível em: <http://pensando.mj.gov.br/dadospessoais/texto-em-debate/anteprojeto-de-lei-para-a-protecao-de-dados-pessoais/> Acesso em: 18 nov. 2016.

BRASIL. Constituição (1988). **Constituição da República Federativa do Brasil**. Brasília DF, 1988. Disponível em: <https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm> Acesso em: 15 out. 2016.

BRASIL. Lei Nº 12.527, de 18 de novembro de 2011. **Diário Oficial da União**, Brasília, 18 nov. 2011, seção 1, p. 1., ed. extra.

BRASSCOM. **Contribuição para consulta pública sobre o Anteprojeto de Lei de Proteção de Dados Pessoais.** São Paulo, SP, 2015. Disponível em: <http://pensando.mj.gov.br/dadospessoais/wp-content/uploads/sites/3/2015/07/19fd0e93b81b716c7e575975aa70b825.pdf> Acesso em: 16 nov. 2016.

CÂMARA DOS DEPUTADOS. **Ficha de tramitação do Projeto de Lei nº 4060/2012.** Brasília, 2012. Disponível em: <http://www.camara.gov.br/proposicoesWeb/fichadetramitacao?idProposicao=548066>
Acesso em: 16 nov. 2016.

CÂMARA DOS DEPUTADOS. **Ficha de tramitação do Projeto de Lei nº 5276/2016.** Brasília, 2016. Disponível em: <http://www.camara.gov.br/proposicoesWeb/fichadetramitacao?idProposicao=2084378> Acesso em: 16 nov. 2016.

CENTRO DE TECNOLOGIA E SOCIEDADE DA FGV DIREITO/RIO**. Contribuição do Centro de Tecnologia e Sociedade da FGV DIREITO RIO ao debate público sobre o Anteprojeto de Lei de Proteção de Dados Pessoais**. Rio de Janeiro, RJ, 2015. Disponível em: <http://pensando.mj.gov.br/dadospessoais/wp-content/uploads/sites/3/2015/07/5c5fb198bc34294eb44cc88dab6a0706.pdf> Acesso em: 16 nov. 2016.

CONDEIXA, F. Direito de privacidade no Brasil. **Revista Jus Navigandi**, Teresina, ano 20, n. 4335, 15 maio 2015. Disponível em: <https://jus.com.br/artigos/33093> Acesso em: 14 nov. 2016.

EL EMAN, K. Has there been a failure of anonymization? *In:* **Electronic Health Information and Privacy**. Disponível em: <http://ehip.blogs.com/ehip/2009/08/has-there-been-a-failure-of-anonymization.html> Acesso em: 05 nov. 2016.

ESTADOS UNIDOS DA AMÉRICA. Other requirements relating to uses and disclosures of protected health information. **Code of Federal Regulations**. Título 45, Seção 164.514. Disponível em: <https://www.law.cornell.edu/cfr/text/45/164.514> Acesso em: 30 out. 2016.

GREENLEAF, G. Global Data Privacy Laws 2015: 109 Countries, with European Laws Now a Minority. **Privacy Laws & Business International Report**, ed.133, p. 14-17. UNSW Law Research Paper No. 2015-21. Sydney, Austrália, 2015. Disponível em: <https://ssrn.com/abstract=2603529> Acesso em: 04 nov. 2016.

GRUPO DE PESQUISA EM POLÍTICAS PÚBLICAS PARA O ACESSO À INFORMAÇÃO DA UNIVERSIDADE DE SÃO PAULO - GPOPAI USP. **Contribuições à Consulta Pública do Anteprojeto de Lei/APL de Proteção de Dados Pessoais**. São Paulo, SP, 2015. Disponível em: <http://pensando.mj.gov.br/dadospessoais/wp-content/uploads/sites/3/2015/07/07c449c076fabbb00f3d3b850e063417.pdf>

IMDB.COM INC. **IMDb Database Statistics**. Disponível em: <http://www.imdb.com/stats> Acesso em: 06 nov. 2016.

ITI - INFORMATION TECHNOLOGY INDUSTRY COUNCIL. **Contribuição ao debate sobre o Anteprojeto de Lei de Proteção de Dados Pessoais.** Washington, EUA, 2015. Disponível em: <http://pensando.mj.gov.br/dadospessoais/wp-content/uploads/sites/3/2015/04/c3c375d635d4870ac5f8228959c69789.pdf> Acesso em: 20 nov. 2016.

MINISTÉRIO DA JUSTIÇA. **Anteprojeto de Lei para a Proteção de Dados Pessoais.** Disponível em: <http://pensando.mj.gov.br/dadospessoais/texto-em-debate/anteprojeto-de-lei-para-a-protecao-de-dados-pessoais/> Acesso em: 15 nov. 2016.

MONTI, M. **Projeto de Lei nº 4060, de 2012.** Dispõe sobre o tratamento de dados pessoais, e dá outras providências. Disponível em: <http://www.camara.gov.br/proposicoesWeb/prop_mostrarintegra?codteor=1001750&filename=PL+4060/2012> Acesso em: 16 nov. 2016.

NARAYANAN, A.; SHMATIKOV, V. Robust De-anonymization of Large Datasets. **SP '08 Proceedings of the 2008 IEEE Symposium on Security and Privacy.** p. 111-125. Washington, EUA, 2008. Disponível em: <https://www.cs.cornell.edu/~shmat/shmat_oak08netflix.pdf> Acesso em: 09 set. 2016.

NARAYANAN, A; SHMATIKOV, V. Myths and fallacies of personally identifiable information. **Communications of the ACM**, v. 53, n. 6, p. 24-26, Nova Iorque, EUA, 2010. Disponível em: <https://www.cs.utexas.edu/~shmat/shmat_cacm10.pdf> Acesso em: 20 set. 2016.

NETFLIX INC. **The Netflix Prize Rules.** 2006. Disponível em: <http://www.netflixprize.com/rules.html> Acesso em: 06 nov. 2016.

OHM, P. Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. **UCLA Law Review**, Vol. 57, p. 1701-1770, U of Colorado Law Legal

Studies Research Paper No. 9-12. Boulder, EUA, 2010. Disponível em: <https://ssrn.com/abstract=1450006> Acesso em: 09 set. 2016.

ORGANIZAÇÃO DAS NAÇÕES UNIDAS. Declaração Universal dos Direitos Humanos. Comunicação & Educação, Brasil, v. 1, n. 3, 2008. Disponível em <http://200.144.189.42/ojs/index.php/comeduc/article/view/4250> Acesso em: 16 nov. 2016.

PORTO SEGURO SERVIÇOS E COMÉRCIO S.A. **Política de Privacidade.** Disponível em: <http://www.portosegurofaz.com.br/politica-de-privacidade> Acesso em: 03 nov. 2016.

PODER EXECUTIVO. **Projeto de Lei nº 5276, de 2016.** Dispõe sobre o tratamento de dados pessoais para a garantia do livre desenvolvimento da personalidade e da dignidade da pessoa natural. Disponível em: <http://www.camara.gov.br/proposicoesWeb/prop_mostrarintegra;jsessionid=5E2716CB19186F756FF72F614115EAC1.proposicoesWebExterno2?codteor=1457459&filename=PL+5276/2016> Acesso em: 16 nov. 2016.

RELX GROUP. **Contribuições à consulta pública sobre o anteprojeto de lei para a proteção de dados pessoais**. 2015. Disponível em: <http://pensando.mj.gov.br/dadospessoais/wp-content/uploads/sites/3/2015/07/d15e68175fc22ca6cda94d3b371f9daa.pdf> Acesso em: 20 nov. 2016.

SWEENEY, L. Achieving k-anonymity privacy protection using generalization and suppression. **International J. of Uncertainty, Fuzziness and Knowledge-based Systems**, Vol. 10, Edição 5, p. 571-578. River Edge, EUA, 2002. Disponível em: <http://dataprivacylab.org/dataprivacy/projects/kanonymity/kanonymity2.pdf> Acesso em: 09 set. 2016.

SWEENEY, L. **Simple demographics often identify people uniquely**. Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh, EUA, 2000. Disponível em: <http://dataprivacylab.org/projects/identifiability/paper1.pdf>. Acesso em: 20 set. 2016.

UNIÃO EUROPEIA - Directiva 95/46/CE do Parlamento Europeu e do Conselho, de 24 de outubro de 1995, relativa à protecção das pessoas singulares no que diz respeito ao tratamento de dados pessoais e à livre circulação desses dados. **Jornal Oficial nº L 281 de 23/11/1995.** p. 31-50. Disponível em: <http://eur-lex.europa.eu/legal-content/pt/TXT/?uri=CELEX%3A31995L0046> Acesso em: 30 out. 2016.

VARON, J. **Contribuição em comentário sobre o inciso XIV do artigo 5 do texto do Anteprojeto de Lei de Proteção de Dados Pessoais.** 2015. Disponível em: <http://pensando.mj.gov.br/dadospessoais/texto-em-debate/anteprojeto-de-lei-para-a-protecao-de-dados-pessoais/> Acesso em 20 nov. 2016.

YAHOO! DO BRASIL INTERNET LTDA. - **Armazenamento de dados e anonimização**. Disponível em: <https://policies.yahoo.com/br/pt/yahoo/privacy/topics/datastorage/index.htm> Acesso em: 03 nov. 2016.

GLOSSÁRIO

Adversário: Entidade, bem ou mal-intencionada, que busca violar a confidencialidade de uma informação.

k-Anonimidade (k-Anonymity): É um modelo de anonimização de dados. Um banco de dados atende ao critério de k-anonimidade se, e apenas se, cada combinação de valores semi-identificadores que ocorra no conjunto de dados ocorrer no mínimo um número “k” de vezes. O valor mínimo para “k” em que a privacidade do banco de dados é preservada é k=2.

Anonimização (Anonymization): Processo que consiste na retirada de informações pessoais de um banco de dados de modo a impedir que os indivíduos a quem os dados se referiam possam ser identificados.

Atributo: Representa uma categoria de informações que consiste em um conjunto de possíveis valores [SWEENEY, 2002, p. 6].

Conjunto de dados de alta dimensão (High dimensional dataset): É um conjunto de dados em que há um grande número de informações sobre cada indivíduo.

Informação Pessoal Identificável (Personally Identifiable Information - PII): É composta por identificadores pessoais e semi-identificadores. Refere-se a todo tipo de informação que pode ser utilizado para identificar, localizar ou contatar uma pessoa.

Identificador (Personal identifier): São atributos que permitem a identificação de um único indivíduo, sem necessidade de informação adicional. Exemplos: CPF, RG, nome completo. Também estão incluídos nesta classificação atributos que permitem uma comunicação direta com o indivíduo, como número de telefone e endereço de e-mail.

INNER JOIN: É uma cláusula presente em Linguagens de Consulta Estruturada (Structured Query Language – SQL) que determina a junção de duas ou mais tabelas

mediante seus valores de atributos em comum. O resultado da operação contém apenas registros presentes em ambas as tabelas.

Reidentificação (Re-identification): É o processo de reversão da anonimização de uma base de dados.

Semi-identificador (Quasi-identifier): São atributos que, em combinação com outros atributos, podem ser associados a um único indivíduo ou um número bastante reduzido de indivíduos.