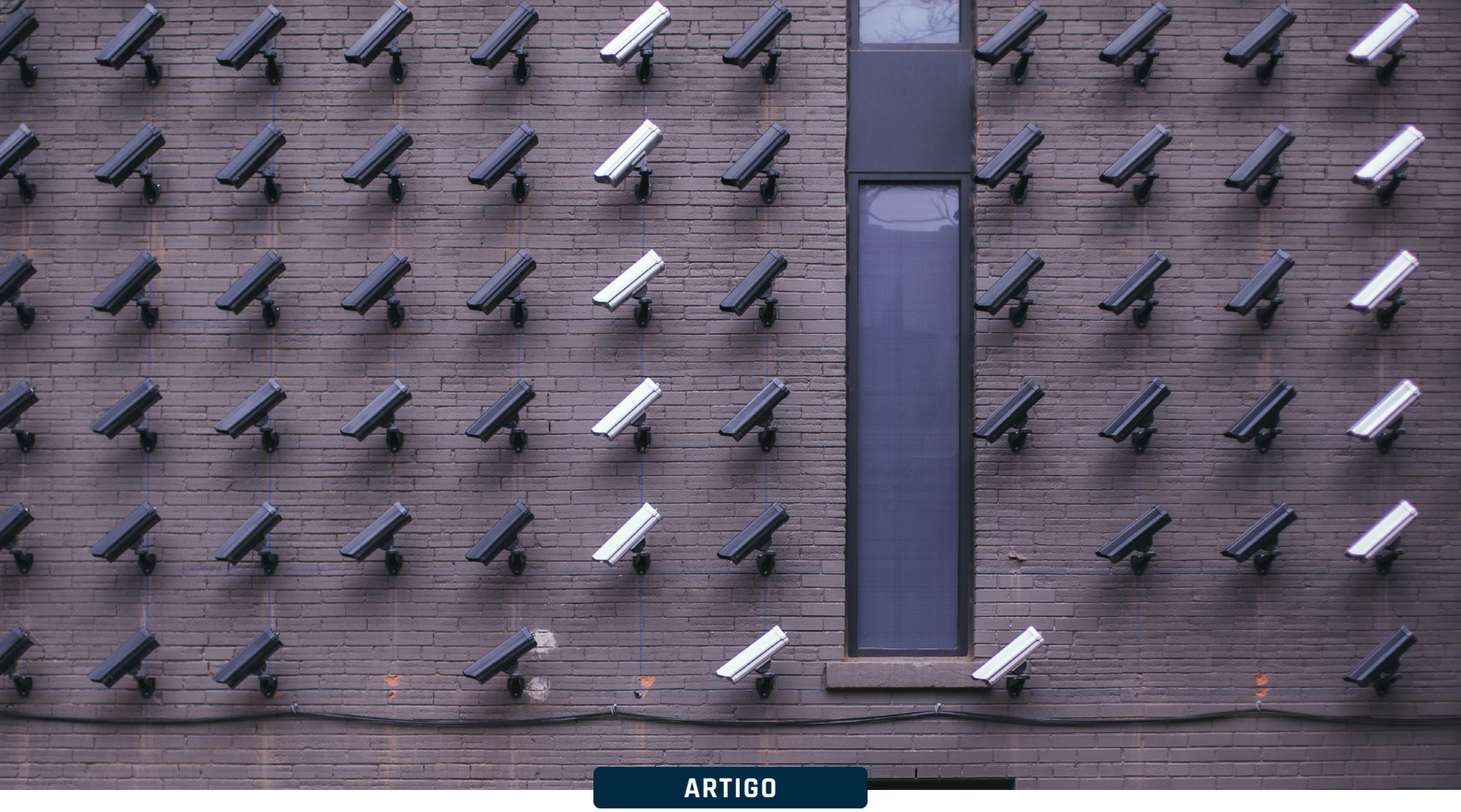

ARTIGO

É POSSÍVEL APRENDER SOBRE AS PESSOAS SEM LHES ESCANCARAR A PRIVACIDADE?

POR

Javam Machado

javam.machado@dc.ufc.br

Organizações modernas - públicas ou privadas - têm continuamente coletado dados pessoais. Quando fazemos buscas na internet, nossos interesses são capturados pelo provedor do serviço de busca. Ao fornecer frequentemente o CPF para acumular pontos em um programa de relacionamento de supermercado, da farmácia de preferência ou do posto de gasolina mais próximo, o indivíduo permite associar padrões de consumo à sua pessoa, e assim facilita a construção de seu perfil histórico de consumo. O hábito crescente de fazer compras em grandes varejistas de vendas online por aplicativos ou sítios web possibilita o mapeamento de interesses individuais e o acúmulo de considerável quantidade de informação sobre as pessoas. Da mesma forma, quando uma pessoa usa um aplicativo de serviço de localização, *streaming* de áudio e de vídeo e redes sociais em seu *smartphone*, ela fornece importante volume de informação pessoal para os provedores desses serviços, seja diretamente por meio das suas escolhas ou informações que ela deliberadamente publica, seja pela coleta de dados que esses aplicativos fazem quando se faz uso de outros aplicativos no mesmo *smartphone*. Os deslocamentos do cidadão nas grandes cidades por meio dos serviços de compartilhamento de bicicle-

tas e de carros elétricos é igualmente uma relevante fonte de informação sobre os indivíduos que fazem uso desse tipo de serviço [6].

A despeito de eventuais abusos na coleta de informação sobre os indivíduos, há o consenso de que a aprendizagem sobre o consumo de bens e serviços, sobre o deslocamento nas cidades, buscas na internet e recomendação de acomodações, áudio e vídeo, todas essas coisas podem servir para melhorar sobremaneira a prestação de serviço, otimizar a venda e a distribuição de bens, definir políticas públicas, enfim facilitar a vida moderna das pessoas. Todavia a aprendizagem, a classificação e a identificação de padrões e de tendências podem ser todas realizadas quando se tem acesso a grandes conjuntos de dados, sem, contudo, representar ou re-identificar o indivíduo que contribui com os seus dados pessoais. Sem que seja possível associar o indivíduo a um item do conjunto de dados, é viável oferecer algum grau de privacidade, assim as pessoas estariam mais seguras para contribuir com o coletivo e não ser importunadas por campanhas de marketing direto, por controle de organizações públicas ou privadas, ou por informações tendenciosas que levam a escolhas induzidas. Desta forma, o que é de consenso pode ser atendido enquanto a privacidade das pessoas é minimamente respeitada.

A Lei Geral de Proteção de Dados (LGPD) discrimina dados pessoais e disciplina o tratamento desses dados pelas organizações a fim de dar garantias de privacidade aos indivíduos [2]. Logo no seu Art 2o inciso I, a LGPD afirma que a disciplina da proteção de dados pessoais tem como fundamento o respeito à privacidade. No seu Art. 5o, inciso I, a mesma norma também define dado pessoal como a informação relacionada à pessoa natural identificada ou identificável. Ainda no Art 5o, agora nos incisos III e XI respectivamente, a LGPD define dado anonimizado como sendo o dado relativo ao titular que não possa ser identificado; e anonimização como a utilização de meios técnicos razoáveis e disponíveis no momento do tratamento, por meio dos quais um dado perde a possibilidade de associação, direta ou indireta, a um indivíduo. A lei, portanto, não só associa dados pessoais ao conceito de privacidade dos indivíduos, como também estabelece o uso de técnicas de tratamento de dados que impossibilitem a associação dos dados aos seus respectivos titulares, ressalvado o uso para objetivos de segurança, combate à criminalidade e jornalismo, dentre outros assemelhados. Mesmo para uso acadêmico, a LGPD determina em seu Art 7o, inciso IV, que esforços sejam feitos para anonimizar os dados pessoais.

A prática comum de anonimização corrente nas organizações que coletam dados pessoais, definidas na LGPD por controladores, procura substituir os identificadores dos indivíduos por valores artificiais, normalmente chamados de pseudônimos. Assim, ao liberar dados para o público ou compartilhar com parceiros, o controlador muitas vezes substitui CPF e nome por valores resultantes de algoritmos que mapeiam os dados

originais para dados fictícios gerados por funções do tipo hash. Entretanto estudos mostram que esse processo de anonimização é insuficiente para impossibilitar a re-identificação do titular do dado ou seja é incapaz de assegurar que o dado público perde a possibilidade de associação, direta ou indireta, a um indivíduo, como preconiza o Art 5º da LGPD [2].

As estratégias mais promissoras para assegurar a perda de associação do titular ao dado publicado buscam associar técnicas de modificação dos dados originais para gerar um conjunto de dados de publicação que mantém características semelhantes ao original, todavia os dados reais dos titulares não estão completamente representados. As principais técnicas de anonimização são a supressão, a generalização e a perturbação [1].

A técnica de supressão de dados remove valores ou substitui um ou mais valores de um conjunto de dados por algum valor especial, impossibilitando a descoberta dos valores originais por um eventual adversário. A generalização aumenta a incerteza de um adversário ao tentar associar um indivíduo a seus dados. Nessa técnica, os valores dos atributos são automaticamente substituídos por valores semanticamente similares, porém menos específicos. A técnica de perturbação substitui os valores dos atributos originais por valores fictícios, mas semelhantes, de modo que informações estatísticas calculadas a partir dos dados originais não se diferenciam significativamente de informações estatísticas calculadas sobre os dados perturbados. Ao contrário das técnicas de generalização e de supressão, que preservam a veracidade dos dados, a perturbação resulta em um conjunto de dados com valores muitas vezes sintéticos [3].

As técnicas de anonimização descritas até aqui são chamadas sintáticas em oposição às técnicas probabilísticas, dentre elas a mais promitente é a privacidade diferencial [4], que goza de fortes garantias de privacidade para os titulares ancoradas em seu arcabouço de definição formal. A privacidade diferencial pode ser vista como um *middleware* entre um estudioso dos dados e um banco de dados, que responde, de maneira privada, a consultas executadas no banco de dados. O *middleware* que implementa a privacidade diferencial é comumente chamado de mecanismo.

Na privacidade diferencial, o mecanismo é dito aleatório porque a probabilidade de qualquer saída desse mecanismo não varia muito com a presença ou ausência de qualquer titular na base de dados [5]. Essa propriedade é garantida pela adição de ruído aleatório controlado à saída da consulta. Normalmente o controle do ruído segue uma distribuição de probabilidade que se assemelha à distribuição de probabilidade da resposta da consulta ao dado original. Assim, apesar de retornar respostas provavelmente fictícias, essa técnica assegura alta dificuldade de re-identificação do titular, enquanto fornece capacidade de análise e estudo do conjunto de dados consultado.

Neste ponto reunimos elementos para responder afirmativamente à pergunta no início deste texto. Entretanto, temos visto o avanço crescente do processo de coleta de dados, construção de perfis

individuais e aprendizagem sobre as pessoas, sem a contrapartida necessária para lhes assegurar a privacidade, a individualidade, o direito à autonomia de escolhas. Precisamos de ferramentas de anonimização formalmente sólidas e disponíveis para integrar esse processo, fortalecendo a aprendizagem e protegendo os indivíduos. A LGPD vai nesse sentido, mas é apenas o passo inicial.

Referências:

1.BRITO, F.; MACHADO, J. Preservação de privacidade de dados: Fundamentos, técnicas e aplicações. 36a JAI - Jornada de Atualização em Informática, Cap 3, pag 1–40. SBC, Porto Alegre, 2017.

2.Congresso Nacional. Lei Geral de Proteção de Dados - Lei No 13.709 de 14 de Agosto de 2018.

3.DOMINGO-FERRER, J.; SANCHEZ, D.; SORIA-COMAS, J. Database anonymization: Privacy, utility, and microaggregation-based inter-model connections. Synthesis Lectures on Information Security, Privacy, and Trust. Morgan & Claypool, 2016

4.DWORK, C. Differential privacy. 33rd International Colloquium on Automata, Languages and Programming, pages 1–12, 2006

5.DWORK, C.; ROTH, A. The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3-4):211–407, 2014.

6.MACHADO, J; DUARTE NETO, E. Privacidade de dados de localização: Modelos, técnicas e mecanismos. 40a JAI - Jornada de Atualização em Informática, Cap 3, pag 105–148. SBC, 2021.

JAVAM MACHADO é professor titular do Departamento de Computação da UFC, onde fundou e coordena o Laboratório de Sistemas e Bancos de Dados (LSBD). Javam foi coordenador da Comissão Especial de Bancos de Dados da SBC (2017) e pesquisador visitante na Telecom SudParis – FR (2001) e no AT&T Labs-Research – USA (2018 e 2020). No momento, o professor Javam se interessa cientificamente pelas áreas de privacidade de dados e de não-discriminação em técnicas de aprendizagem automática.