Aspectos computacionais do tratamento de dados pessoais no âmbito da Lei Geral de Proteção de Dados Pessoais

Evandro Eduardo Seron Ruiz*
Departmento de Computação e Matemática
FFCLRP – USP
Ribeirão Preto, SP. Brasil
evandro@usp.br

June 28, 2019

Abstract

A nova Lei Geral de Proteção de Dados Pessoais que entrará em vigor a partir de agosto de 2020 deverá mudar radicalmente a forma de tratamento dos dados pessoais dos brasileiros. As garantias fundamentais de privacidade e liberdade quanto a este tipo de dado estão intrinsecamente ligadas às capacidades de métodos computacionais que promoverão o anonimato destes dados. Este projeto tem como objetivo produzir materiais técnico-científicos sobre metodologias de desidentificação, anonimização, desanonimização e pseudoanonimização de dados pessoais, principalmente os dados abertos divulgados pelas administrações públicas de organizações municipais, estaduais e federais.

privacidade anonimização reidentificação

1 Introdução

Os serviços de armazenamento de dados em nuvem parecem hoje serem infindáveis. Praticamente todo e qualquer tipo de informação parece estar ao alcance dos dedos quando em contato com um computador ou telefone pessoal. Se estes dados abertos ao público podem oferecer inúmeros benefícios à sociedade em geral, às pessoas e às organizações, a eventual reutilização destes dados deve respeitar o direito de todos à privacidade e a proteção de seus dados pessoais.

Atualmente milhões de dados pessoais tais como nome, filiação, endereço e números de documentos circulam pela web. São dados transferidos por email, redes sociais, postados na

*Mestre pela UNICAMP, Ph.D. pela University of Kent at Canterbury, Inglaterra, Prof. Livre-Docente pela USP, Professor Associado, DCM – FFCLRP, USP.

web e até em sites oficias. Imagine agora que uma pessoa vá até um laboratório de análises clínicas para realizar alguns exames biomédicos. Hoje a posse dos dados, resultados destes exames, são de responsabilidade do laboratório que oferece aos seu clientes uma senha para recuperá-los, da sua base de dados, usando recursos da web. Um cliente, de posse de seu telefone pessoal, um *smartphone*, acessa estes dados, armazena-os no seu telefone e depois os envia para seu médico. Todas essas informações são trocadas através de pacotes de dados na web que circulam por redes proprietárias, algumas destas podem ser redes seguras, protegidas por senhas e protocolos de segurança, mas outras podem ser redes abertas, desprotegidas. O envio dos dados pode ocorrer por meio de serviço de email ou pelo uso de um comunicador de mensagens, um mensageiro eletrônico talvez. O médico, por sua vez, pode receber estas informações no seu computador pessoal, no seu telefone, ou mesmo usar um computador de uso comum no seu ambiente de trabalho. Neste comutador ele pode armazenar os dados do paciente para poder lê-los. Pronto! Este é um cenário comum e perfeito para analisarmos brechas de segurança na transação e no armazenamento de dados pessoais.

Dadas as situações cotidianas, como a citada anteiormente, podemos analisar várias passagens que demonstram as potenciais vulnerabilidades no tratamento de dados pessoais, tanto nas formas de armazenamento quanto na transmissão destes dados. Não é difícil imaginar que situações semelhantes ocorram na troca de mensagens e documentos entre escritórios contábeis e seus clientes, agências de seguro, agências de viagens entre outras.

No entanto, estes dados poderiam ser anonimizados, os nomes trocados por códigos que somente os geradores destes códigos pudessem re-identificálos. Sendo assim, teríamos cenas perfeitas de proteção de dados para os que realmente dela fazem uso esperado e cenas de desalento para quem pretendia usar estes dados com intenções duvidosas. Ainda assim, parafraseando Ohm (2010), "Dados não podem ser, ao mesmo tempo, úteis e perfeitamente anônimos". O que Ohm (2010) afirma na frase acima é que existem técnicas otimizadas que asseguram elevadas taxas de reidentificação ou desanonimização de dados pessoais mas dados anonimizados podem ter pouca utilidade para a sociedade como um todo.

Esses processos, ou mesmo, essas técnicas de anonimização e reidentificação não só desafiam a recente Lei Geral de Proteção de Dados Pessoais (CIVIL, 2018), como abrem um grande debate na sociedade sobre aspectos de proteção, privacidade, confidencialidade e a real utilidade destes tipos de dados.

A Lei Geral de Proteção de Dados Pessoais (CIVIL, 2018) (LGPDP), sancionada pela Presidência em 14 agosto de 2018, deverá entrar em vigor a partir de agosto de 2020. Esta nova lei, em seu texto determina que todos os dados pessoais só podem ser coletados mediante o consentimento livre e esclarecido do usuário. Determina também que não se deve realizar o tratamento destes dados pessoais sem o consentimento do titular. A LGPDP deverá disciplinar o tratamento dos dados pessoais, tais como, seu nome e sobrenome, CPF e RG, além de dados como sua etnia, religião, sexualidade e opinião política que são tidos como dados sensíveis e

merecedores proteção. Assim sendo, espera-se que esta nova lei deva transformar radicalmente o modo como nossos dados pessoais são tratados por algumas organizações, empresas e até pessoas físicas no Brasil.

Cabe ressaltar que a LGPDP encontra similaridades e orientações na *General Data Protection Regulation*, GDPR, da União Europeia, onde é obrigatória desde maio de 2018 também com o *California Consumer Privacy Act of 2018*, CCPA, dos Estados Unidos da América, onde foi aprovado em junho de 2018

A LGPDP é ainda mais abrangente quando prevê a criação da Autoridade Nacional de Proteção de Dados, ANPD, criada pela medida provisória 869, de 27 de dezembro de 2018, a quem caberá a determinação de padrões técnicos mínimos para viabilizar as medidas de segurança, além das medidas técnicas e administrativas aptas a proteger os dados pessoais de acessos e processamentos ilícitos. Ressaltamos que até este momento a ANPD ainda não tem a sua composição definida e, mais importante, ainda não temos as bases técnicas para a implementação desta lei.

Sob o ponto de vista da Ciência da Computação, os processos de desidentificação de dados, anonimização, desanonimização e pseudoanonimização são processos intrinsecamente técnicos que não podem prescindir de análises de aplicabilidade de seus algoritmos, da robustez das tecnologias auxiliares usadas e da avaliação de e erros típicos decorrentes do uso deste tipo de tecnologia em ampla escala. Ressalto que dada a novidade destas diversas leis sobre a proteção de dados pessoais, em âmbito nacional e internacional, ainda existe um longo caminho para a viabilização técnica de métodos computacionais que permitam respeitar e manter a privacidade de dados pessoais.

A descrição deste projeto está de acordo com a sequência de seções sugeridas para a criação de projetos do *Programa Ano Sabático* do Instituto de Estudos Avançados, IEA, da Universidade de São Paulo.

2 Objetivo

Este projeto tem como objetivo principal produzir materiais técnico-científicos sobre metodologias de desidentificação, anonimização, desanonimização e pseudoanonimização de dados pessoais, principalmente os dados abertos divulgados pelas administrações públicas de organizações municipais, estaduais e federais. Esperamos que os artigos científicos os os materiais de divulgação eventualmente produzidos sejam úteis para esclarecer e alinhar as atividades da Autoridade Nacional de Proteção de Dados e de outras agências, a exemplo da Unidade Especial de Proteção de Dados e Inteligência Artificial do Ministério Público do Distrito Federal.

2.1 Objetivos secundários

São objetivos secundários:

1. Recuperar e fazer uma síntese técnica, uma compilação, de todos os métodos de tratamento de dados para desidentificação, anonimização, desanonimização e pseudoanonimização referenciados pelas demais autoridades de proteção de dados de outros países, tais como: Áustria, Bélgica, Dinamarca, Croácia, Uruguai, Nova Zelândia, EUA, entre outros;

2. Compilar algumas fontes públicas de dados abertos e pessoais, entre estes, dados de saúde, os quais possuem registros não anonimizados ou com potencial desanonimização trivial;

3. Estudar a aplicação dos principais métodos de desidentificação, anonimização, desanonimização e pseudoanonimização sobre dados públicos abertos.

Pretendo também produzir materiais técnicos de divulgação sobre os resultados obtidos em cada uma destas fases.

3 Justificativa

A LGPDP ajuda a formatar um novo arcabouço legal para a proteção de dados pessoais no Brasil. Sua principal função é ordenar e alinhar os direcionamentos práticos da futura ANPD. O objetivo de ambos é "proteger os direitos fundamentais de liberdade e de privacidade e o livre desenvolvimento da personalidade da pessoa natural" (CIVIL, 2018). As principais 'ferramentas' garantidoras da LGPDP são os algoritmos de desidentificação, anonimização, desanonimização e pseudoanonimização de dados pessoais. Assim o estudo destas metodologias promovem:

- Manter e preservar do uso ilegal os ativos mais valiosos e mais vulneráveis das pessoas que são justamente os seus dados pessoais;

- Os dados pessoais atualmente são os alvos principais de hackers e algumas empresas que usam indevidamente dados pessoais de forma irrestrita. Os *ransomware* (sequestro de dados tornando alguns dos dados disponíveis no equipamento totalmente inacessíveis) e os esquemas de *phishing* (práticas de tentativas de obtenção de informação pela suplantação de identidade por parte de criminoso);

- A violação de dados pode ocorrer com qualquer pessoa ou instituição. Não devemos considerar o "se acontecer" mas sim "quando vai acontecer";

- As possibilidade de desanonimização crescem a medida que cresce o número e a variabilidade de dados pessoais na web. Isso torna a pesquisa em métodos de desidentificação, anonimização, desanonimização uma prática contínua.

Assim, a pesquisa interdisciplinar nos aspectos computacionais de proteção de dados pessoais justificam-se para manter um dos maiores direitos humanos, o direito a liberdade.

4 Razões para desenvolver o projeto no IEA

Considero este trabalho como pertencente a uma área interdisciplinar, a junção de aspectos de Direito Civil com técnicas e metodologias da Ciência da Computação. Há alguns anos tenho contato com a Profa. Dra. Cíntia Rosa Pereira de Lima, professora associada da FDRP-USP. Este contato propiciou o desenvolvimento de alguns pequenos projetos conjuntos em temas que envolvem Direito e Tecnologias da Informação. Este projeto de pesquisa nasceu desta pareceria e encontramos no Instituto de Estudos Avançados um potencial espaço de integração de áreas de conhecimento para fomentar tanto as nossas pesquisas como eventuais discussões com a comunidade sobre este tema de proteção de dados pessoais.
Destacamos que a missão do IEA de pesquisar, junto com segmentos representativos da sociedade, sobre temas de impacto em políticas públicas, é totalmente aderente aos objetivos deste projeto que procura alicerçar os limites da Computação às garantias de liberdade individual dos cidadãos com uma 'vida' online, através do correto tratamento dos seus dados pessoais.

5 Potencial de interdisciplinaridade

Este é um trabalho que busca interpretar os anseios da LGPDP e as eventuais diretivas práticas da ANPD que garantirá a proteção dos direitos fundamentais de liberdade e de privacidade relativas as dados pessoais da pessoa natural. Pode ser assim considerado um projeto interdisciplinar que procura, com o auxílio de métodos computacionais, garantir o tratamento legal e correto de dados pessoais preservando o seu anonimato.
A Ciência da Computação tem os meios técnicos de investigar, por exemplo, se há garantias da anonimato pelo simples processo de desidentificação de dados pessoais aos quais são submetidos os atuais dados de saúde de todos os usuários dos sistemas de saúde, público ou privado, no território brasileiro. Quanto a outros dados pessoais em poder das mais diversas organizações públicas e privadas, são os métodos computacionais que deverão trabalhar para a anonimização e desanonimização destes dados pessoais para continuar garantindo os direitos fundamentais de proteção sobre estes dados.
Obviamente todo a conceituação de direitos e garantias individuais sobre os dados de cada cidadão deve ser atribuída aos profissionais de Direito Civil. Estas conceituações partem de

pressupostos hoje tido como elementares, tais como: o respeito à privacidade; a autodeterminação informativa; a liberdade de expressão, de informação, de comunicação e de opinião; a inviolabilidade da intimidade, da honra e da imagem, entre outras. Todos este fundamentos que hoje desfrutamos e que são preservados pelos profissionais de Direito. Portanto, cabe aos profissionais da Ciência da Computação garantir a eficácia e os limites das técnicas de anonimização existentes no contexto jurídico comunitário de proteção de dados e apresentar recomendações para lidar com essas técnicas.

6 Impactos científicos e sociais

Os impactos da LGPDP na sociedade brasileira podem ser previstos se, por exemplo, analisarmos aos impactos da sua lei irmã na União Européia, a *General Data Protection Regulation*, GDPR. Como exemplo, Shu e Jahankhani (SHU; JAHANKHANI, 2017) detalham os impactos da GDPR no sistema público de saúde do Reino Unido e a criação da sua autoridade garantidora de dados pessoais, a *National Data Guardian*, NDG. Por outro lado, Stevens em seu artigo (STEVENS, 2015)faz uma análise crítica do impacto da GDPR nas pesquisas em ciências sociais que envolvam dados administrativos dos países da própria UE. Mais recentemente, Butterworth (BUTTERWORTH, 2018) analisou os impactos de uma eventual cultura a termo longo que se poderá se apropriar de métodos de inteligência artificial que poderão, por sua vez, até induzir a novos paradigmas morais e éticos no futuro.

Os trabalhos acadêmicos citados acima são apenas exemplos de como esta nova área interdisciplinar pode motivar a sociedade a discutir eventuais situações provocadas por mudanças no modo de tratamento de dados escolhido pela própria sociedade. Os trabalhos esperados a partir deste projeto deverão focar na análise e recomendação de metodologias para tratamento de dados pessoais mas as consequências da aplicação destes algoritmos devem alterar a visão comum sobre vários conceitos de proteção de dados pessoais.

7 Metodologia

Como já afirmamos na Seção 2, este projeto tem como objetivo principal produzir materiais técnico-científicos sobre metodologias de desidentificação, anonimização, desanonimização e pseudoanonimização de dados pessoais, principalmente os dados abertos divulgados pelas administrações públicas de organizações municipais, estaduais e federais.

7.1 Anonimização

Em termos gerais, existem duas abordagens distintas de anonimização de dados pessoais: a primeira tem por base a aleatorização, enquanto a segunda se baseia na generalização.

Anonimização é uma solução para a remoção de informações sensíveis de um documento. Carvalho Dias, especificamente define anonimização de dados como o nome dado a um processo para mascarar ou remover informações sensíveis de um documento preservando seu formato original (DIAS, 2016). Anonimização também pode ser entendido como o processamento irreversível de dados pessoais de forma não prever a identificação de uma pessoa (GT216, 2014).

Houve uma época em que se acreditava que haveriam meios de promoção da chamada 'anonimização robusta', ou seja, a possibilidade de anonimizar dados de tal maneira que a reidentificação não fosse possível. Uma maneira de anonimização robusta seria a possibilidade de usarmos somente dados anonimizado, o que diminuiria drasticamente a utilidade de muitos desses dados.

Segundo Acs, Castelluccia e Le Métayer (2016) existem três critérios que podem ser usados para avaliar a robustez de um processo de anonimização, que são:

Distinção: ou seja, a possibilidade de isola alguns ou todos os registros que identificam um indivíduo num conjunto de dados;

Capacidade de relacionamento: que é a capacidade de relacionamento, ou de ligação, de ao menos dois registros relacionados a um mesmo indivíduo ou grupo de indivíduos sobre os quais os dados seus dados estão sendo coletados; e

Inferência: que é a possibilidade de deduzir, com razoável grau de probabilidade, o valor de um atributo conhecidos os valores de outros atributos.

Ressalto que os dois primeiros critérios são relativos a inferência de identidade enquanto o terceiro é relativo a inferência de atributos. As inferências de identidade recupera a identidade dos registros anonimizado enquanto as inferências de atributos recuperam características de um registro. El Emam e Álvarez (2014) argumentam que a regulamentação de proteção de dados deve focar nas questões relacionadas a reidentificação, ou seja, devem ter foco na inferência de identificação. A prevenção e o zelo pela inferência de atributos deve ser uma questão para os comitês de ética.

Deveremos pesquisar as seguintes técnicas de anonimização:

7.2 Aleatorização

A aleatorização é uma família de técnicas que altera a veracidade dos dados a fim de eliminar a estreita ligação entre os dados e a pessoa. Se os dados forem suficientemente imprecisos já não poderão ser relacionados com uma pessoa específica. A aleatorização não reduz, por si só, a singularidade de cada registo, uma vez que cada registo continua a ser proveniente de um único titular dos dados, mas é passível de proteger contra ataques ou riscos de inferência e pode ser combinada com técnicas de generalização a fim de fornecer garantias de privacidade mais sólidas.

7.3 Adição de ruído

A técnica de adição de ruído é especialmente útil quando os atributos são passíveis de ter um grande efeito adverso sobre as pessoas e consiste em modificar atributos no conjunto de dados de modo a este serem menos precisos, enquanto se mantém a distribuição global. Ao efetuar o tratamento de um conjunto de dados, um observador irá presumir que os valores são exatos, o que só será verdade até um certo nível. Por exemplo, se a altura de uma pessoa tiver sido originalmente medida até ao centímetro mais próximo, o conjunto de dados anonimizados pode conter uma altura com uma precisão arredondada ao intervalo de 10 cm mais próximo. Se esta técnica for aplicada eficazmente, um terceiro não conseguirá identificar uma determinada pessoa, nem tão pouco conseguirá reparar os dados ou detetar de que forma estes foram alterados. Frequentemente a adição de ruído necessita de ser combinada com outras técnicas de anonimização de dados pessoais, tais como a remoção de atributos evidentes e de quase-identificadores.

7.4 Privacidade diferencial

A privacidade diferencial (DWORK, 2006) inclui-se na família de técnicas de aleatorização, com uma abordagem diferente: enquanto, na verdade, a inserção de ruído se aplica previamente à divulgação do conjunto de dados, a privacidade diferencial é passível de ser utilizada quando o responsável pelo tratamento de dados gera visualizações anonimizadas de um conjunto de dados, conservando uma cópia dos dados originais. Essas visualizações anonimizadas seriam normalmente geradas através de um subconjunto de consultas para um terceiro em especial.

7.5 Agregação e k-anonimato

As técnicas de agregação (SHI et al., 2011) e k-anonimato (POPA et al., 2011) visam impedir que um titular dos dados seja selecionado através do agrupamento com, pelo menos, outras k pessoas. Para este efeito, os valores dos atributos são generalizados de modo a que cada pessoa partilhe o mesmo valor. Por exemplo, ao reduzir a granularidade de um local de uma cidade para um país é incluído um maior número de pessoas. As datas de nascimento individuais podem ser generalizadas num intervalo de datas ou agrupadas por mês ou ano. Outros atributos numéricos (por exemplo, salários, peso, altura ou a dosagem de um medicamento) podem ser generalizados por intervalos de valores. Estes métodos podem ser utilizados quando a correlação entre valores pontuais de atributos é passível de criar quase-identificadores.

Pretendemos também discutir como estas técnicas de anonimização são seguras em relação as técnicas de desanonimização (DING et al., 2010).

8 Plano de trabalho

O plano de trabalho está dividido em três fases que respondem aos objetivos secundários. São estas:

Fase 1 Recuperar e fazer uma síntese técnica, uma compilação, de todos os métodos de tratamento de dados para desidentificação, anonimização, desanonimização e pseudoanonimização referenciados pelas demais autoridades de proteção de dados de outros países, tais como: Áustria, Bélgica, Dinamarca, Croácia, Uruguai, Nova Zelândia, EUA, entre outros;

Fase 2 Compilar algumas fontes públicas de dados abertos e pessoais, entre estes, dados de saúde, os quais possuem registros não anonimizados ou com potencial desanonimização trivial;

Fase 3 Estudar a aplicação dos principais métodos de desidentificação, anonimização, desanonimização e pseudoanonimização sobre dados públicos abertos.

9 Cronograma

Como cronograma, pretendemos desenvolver as três fases descritas na Seção 8 acima nos seis meses propostos no projeto mais um períodos dedicados a publicação de resultados. Sendo assim, este é o planejamento:

Fase 1 A ser realizada entre os meses de fevereiro e março;

Fase 2 Entre os meses de março e abril, e, finalmente;

Fase 3 Nos meses de maio e junho.

10 Elaboração de trabalhos científicos

Espero que o trabalho desenvolvido seja o suficiente para a publicação de artigos científicos em periódicos indexados e com corpo editorial. Julgo importante a veiculação em periódicos internacionais e também, dada as características precursoras da pesquisa, crio que seja importante a divulgação dos trabalhos também em língua portuguesa. Pelo planejamento proposto penso que há possibilidade de publicação de dois artigos acadêmicos sobre o tema explorando o atual cenário brasileiro de exposição de dados públicos e as sugestões metodológicas para a Autoridade Nacional.

11 Previsão de organização de seminários ou simpósios

No escopo deste trabalho conjunto com a Profa. Cíntia Rosa Pereira de Lima da FDRP, deverei participar como membro de uma das mesas de discussão do Congresso Internacional "Desafios e novas perspectivas sobre as Autoridades de Proteção de Dados Pessoais e Privacidade", a ser realizado no Auditório da FDRP no período dos dias 07 a 09 de novembro de 2019. Estamos também colocando nossos estudantes de iniciação científica e pós-graduação em contato para promoção de seminários quinzenais em Direito e Tecnologia da Informação. Esta parceiria acadêmica esta propondo também uma nova disciplina optativa para estudantes dos cursos de graduação em Direito e Ciência da Computação chamada de "Direito, Tecnologia da Informação e Web". Estas três iniciativas devem ser abertas a sociedade, mais especificamente, pretendemos derivar alguns seminário de divulgação de caráter mais técnico ainda no primeiro semestre de 2020 sobre os temas aqui estudados.

References

ACS, G.; CASTELLUCCIA, C.; Le Métayer, D. Testing the robustness of anonymization techniques: acceptable versus unacceptable inferences. In: *The Brussels Privacy Symposium*. [S.l.: s.n.], 2016. p. 1–7. ⟨https://fpf.org/wp-content/uploads/2016/11/Acs_CL-DPL16-v4.pdf⟩.

BUTTERWORTH, M. The ico and artificial intelligence: The role of fairness in the gdpr framework. *Computer Law & Security Review*, Elsevier, v. 34, n. 2, p. 257–268, 2018.

CIVIL, C. *Lei Geral de Proteção de Dados Pessoais*. 2018. ⟨http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm⟩.

DIAS, F. M. C. *Multilingual Automated Text Anonymization*. 134 p. Tese (Doutorado) — Universidade Técnica de Lisboa, 2016.

DING, X. et al. A brief survey on de-anonymization attacks in online social networks. In: IEEE. *2010 international conference on computational aspects of social networks*. [S.l.], 2010. p. 611–615.

DWORK, C. Differential privacy. automata, languages and programming, pt 2, 4052: 1–12, 2006. bugliesi, m prennel, b sassone, v wegener. In: *I 33rd International Colloquium on Automata, Languages and Programming JUL*. [S.l.: s.n.], 2006. p. 10–14.

El Emam, K.; ÁLVAREZ, C. A critical appraisal of the Article 29 Working Party Opinion 05/2014 on data anonymization techniques. *International Data Privacy Law*, v. 5, n. 1, p. 73–87, 2014. ISSN 2044-3994.

GT216. *Opinion 05 / 2014 on Anonymisation Techniques Adopted. Data Protection Working Party, EU*. 2014. ⟨https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf⟩.

OHM, P. Broken promises of privacy: responding to the surprising failure of anonymization. *UCLA Law Review*, v. 57, p. 1701–1777, 2010.

POPA, R. A. et al. Privacy and accountability for location-based aggregate statistics. In: ACM. *Proceedings of the 18th ACM conference on Computer and communications security*. [S.l.], 2011. p. 653–666.

SHI, E. et al. Privacy-preserving aggregation of time-series data. In: INTERNET SOCIETY. *Annual Network & Distributed System Security Symposium (NDSS)*. [S.l.], 2011.

SHU, I. N.; JAHANKHANI, H. The impact of the new european general data protection regulation (gdpr) on the information governance toolkit in health and social care with special reference to primary care in england. In: IEEE. *2017 Cybersecurity and Cyberforensics Conference (CCC)*. [S.l.], 2017. p. 31–37.

STEVENS, L. The proposed data protection regulation and its potential impact on social sciences research in the uk. *Eur. Data Prot. L. Rev.*, HeinOnline, v. 1, p. 97, 2015.