A privacidade, a segurança da informação e a proteção de dados no *Big Data*

Antonio João Gonçalves de Azambuja[1], *Lisandro Zambenedetti Granville*[2] *e Alexandre Guilherme Motta Sarmento*[3]

Resumo

O avanço das tecnologias da informação tem possibilitado um crescimento exponencial do volume de dados obtidos, armazenados, processados, transmitidos e publicados no ambiente do *Big Data*. Todo esse crescimento tem gerado desafios para o direito à privacidade, à liberdade de expressão e a segurança das informações, tanto pessoais como as corporativas. As questões do volume de dados, a velocidade com que os dados são processados, a sua variedade e veracidade no ecossistema do *Big Data* colocam em risco esses direitos e a segurança das informações. Inicialmente, este trabalho apresenta

Abstract

The progress of information technologies has enabled an exponential growth in the volume of data collect, stored, processed, transmitted and published in the Big Data environment. All of this growth has created challenges for the right to privacy, freedom of expression and the security of both personal and corporate information. Data volume issues, the speed with which data is processed, its variety and veracity in the Big Data ecosystem, put those rights and the security of information in risk. Initially this paper presents a contextualization about Big Data, with definitions and their characteristics. Then

1 Chefe do Serviço de Segurança da Informação e Comunicações da Advocacia-Geral da União. Mestre em Gestão do Conhecimento e Tecnologia da Informação pela Universidade Católica de Brasília (UCB).

2 Professor do Programa de Educação em Ciências Universidade Federal do Rio Grande do Sul (UFRGS). Doutor em Computação pela UFRGS.

3 Coordenador técnico de apoio a pesquisa, desenvolvimento e aplicações do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Doutor em Educação em Ciências pela UFRGS.

uma contextualização sobre o *Big Data*, com definições e suas características. Em seguida aborda questões relacionadas com a ética, a privacidade, a segurança e a organização das informações no *Big Data*. Ao abordar tais questões, discorre sobre os riscos à privacidade, segurança da informação, organização da informação, conceitos e modelos de anonimização que podem ser utilizados para preservar a privacidade dos usuários. Finalmente, apresenta um panorama da proteção de dados em 2018, com os eventos de divulgação e manipulação de dados sem autorização dos usuários, fato que direciona para maior cuidado com os nossos dados. As considerações finais da análise reconhecem que os usuários estão sujeitos aos riscos à privacidade das informações e a sua segurança no universo do *Big Data*.

Palavras-chave: Privacidade das informações. Segurança da informação. Risco à Privacidade. Proteção de dados. *Big Data*.

addresses issues related to ethics, privacy, security and the organization of information in Big Data. Addressing such issues related to privacy risks, information security, information organization, anonymization concepts and models that can be used to preserve users' privacy. Finally presents a panorama of data protection in 2018, the events of disclosure and manipulation of data without users authorization. That fact leads to a greater care with our data. The final considerations in the analysis recognize that users are subject to the risks to information privacy and their security in the Big Data universe.

Keywords: *Privacy information. Information Security. Privacy risk. Data protection. Big Data.*

1. Introdução

A informatização da sociedade, aliada ao avanço tecnológico e a sua convergência, tem proporcionado um crescimento exponencial do volume de dados no espaço cibernético, o que marca o advento do *Big Data*.

Vivemos a era do *Big Data*, que tem transformado a forma como as organizações estão direcionando o seu processo de tomada de decisão (JANSSEN *et al.*, 2017). As novas tecnologias permitem que as organizações, a partir da análise dos dados, tenham um ganho de competitividade (EREVELLES; FUKAWA; SWAYNE, 2016).

Nesse cenário, composto pela explosão da quantidade e disponibilidade de dados decorrente do avanço das tecnologias de processamento, coleta e análise dos dados, situa-se o fenômeno conhecido como *Big Data*.

Esse fenômeno tem impactado a sociedade, por meio de novos modelos de negócios que fazem rastreamento de dados para analisar padrões de comportamento, consumo e saúde, visando a estabelecer uma tomada de decisão baseada em dados.

Ao mesmo tempo que novas formas de comunicação, registro, acesso e recuperação da informação estão sendo viabilizadas, surge a preocupação com a privacidade e segurança das informações (CHEN; YANG; LUO, 2017).

A privacidade e a segurança da informação (SI) na internet têm correspondido a uma área que desperta interesse de estudo, devido à grande quantidade de informações pessoais e corporativas que são obtidas, armazenadas, transmitidas e publicadas na rede mundial de computadores.

A informação tornou-se um ativo de valor para as organizações, que pode ser processada por meio eletrônico e com a utilização de redes públicas e privadas de internet (HONG; THONG, 2013).

Esses ativos que podem ser corporativos e/ou pessoais compõem o ambiente atual dos negócios das organizações e estão em constante ameaça de vírus, invasões de sistema, abuso de informações privilegiadas, quebra da privacidade e divulgação não autorizada das informações (JOHNSTON; WARKENTIN, 2010).

Para Zwitter (2014), em um mundo altamente interconectado, lidar com a ética, considerando o consentimento dos usuários, a privacidade, a segurança e a anonimização das informações, é um desafio do *Big Data*. Já Drinkwater (2016), ressalta que o vazamento de informações *on-line*, aumenta as preocupações dos usuários em relação ao risco da informação.

Diante do contexto no qual os direitos à privacidade e proteção de dados foram elevados ao nível dos direitos humanos no cenário internacional, os governos têm dispensado especial atenção para lidar com esses desafios. Nesse cenário, destaca-se o Regulamento Geral sobre a Proteção de Dados (GDPR), publicado em 2018, pela União Europeia (EU), que visa a proporcionar aos usuários maior controle sobre seus dados pessoais e a aumentar as restrições sobre as organizações que tratam e lidam com esses dados.

No cenário nacional, por sua vez, o Governo Brasileiro publicou a Lei Geral de Proteção de Dados Pessoais (LGPD), n.º 13.709, de 14 de agosto de 2018, que dispõe sobre o tratamento de dados pessoais, inclusive nos meios digitais, por pessoa natural ou pessoa jurídica de direito público ou privado, com o objetivo de proteger os direitos fundamentais de liberdade e de privacidade e o livre desenvolvimento da personalidade da pessoa natural. A referida Lei entrará em vigor no primeiro semestre de 2020.

Diante do exposto, este artigo busca apresentar uma contextualização sobre o *Big Data*, discorrer sobre questões relacionadas com a ética, a privacidade, a segurança e a organização das informações nesse ambiente, apresentar conceitos e modelos de anonimização que podem ser utilizados para preservar a privacidade dos usuários, além de um panorama sobre a proteção de dados em 2018.

2. *Big Data*

2.1. Definição

O *Big Data* é um fenômeno que se refere à explosão da disponibilidade de dados relevantes, como resultado recente e sem precedente do avanço das tecnologias de armazenamento e registro de dados. Fenômeno do processamento de grandes volumes de dados, com os quais as ferramentas tradicionais não são capazes de lidar na velocidade requerida (GOLDMAN *et al.*, 2012).

Brynjolfsson *et al.* (2012) afirmam, ainda, que soluções de *Big Data* possuem um potencial maior que as soluções analíticas tradicionais, no sentido de trazer benefícios e aumentar a competitividade das empresas.

O termo *Big Data* surgiu para definir arquiteturas de sistemas capazes de lidar com as novas dimensões dos dados: velocidade, variedade e volume (AZEVEDO; NEVES; NOVO, 2014).

Nesse cenário de crescimento exponencial da informação publicada na internet, com a presença de base de dados que contém um grande volume de dados, situa-se o *Big Data* (SHINATAKU; DUQUE; SUAIDEN, 2014).

2.2. Características

O fenômeno *Big Data* está associado ao grande volume de dados, mas essa não é sua única característica. Inicialmente, foi caracterizado pelo volume, pela velocidade e variedade (3V´s) dos dados. Os atributos veracidade e valor foram considerados posteriormente como relevantes. Essas características são conhecidas como os 5V´s do *Big Data*:

- *Volume:* refere-se ao tamanho dos dados. Os dados são coletados de uma grande variedade de fontes, incluindo transações comerciais, redes sociais e informações de sensores ou dados transmitidos de máquina a máquina;

- *Velocidade:* refere-se à velocidade de transmissão dos dados. Os dados fluem em uma velocidade sem precedentes e devem ser armazenados, tratados e analisados com agilidade;

- *Variedade:* refere-se ao formato no qual os dados são gerados, isto é, estruturados e não estruturados. Os dados estruturados são organizados em linhas e colunas e geralmente são armazenados em banco de dados relacionais, os quais facilitam a atualização e a recuperação de dados em menor granularidade. Os dados não estruturados não possuem uma organização predefinida. Em decorrência disso, há maior dificuldade para a sua recuperação e o seu processamento, a exemplo dos vídeos, dos comentários em redes sociais, dos e-mails, entre outros;

- *Veracidade:* tem relação com a confiabilidade dos dados. Durante a análise dos dados, é necessário conhecer o contexto em que os dados foram gerados, se eles são autênticos e de fontes confiáveis; e

- *Valor:* os dados devem agregar valor ao negócio. Sem valor, a informação não tem utilidade.

2.3. Fonte

As fontes de dados do *Big Data* são: os usuários e a tecnologia. Dados, informações e conhecimento são gerados diariamente. A complexidade do *Big Data* não está no volume, como apontou Davenport (2014), mas na falta de estrutura que dificulta a análise para a geração de conhecimento, inovação ou valor.

O autor destaca a relevância de se resumir os dados e encontrar seus significados e seus padrões para o contexto no qual ele foi resumido. Reforça a importância da definição adequada do problema e da pertinência da formulação correta da pergunta, os quais devem orientar a coleta e o posterior resumo dos dados, na busca da organização da informação.

3. Ética

Questões éticas devem ser consideradas no atual cenário do *Big Data*, tais como: Qual a fronteira para o uso dos dados produzidos pelas pessoas no seu dia a dia, com as novas tecnologias? Esses dados podem ser acessados em tempo real? Por quem? Para que finalidade?

O volume de dados cresce de forma exponencial com a evolução e sofisticação da rede mundial de computadores e de suas aplicações. Todo o potencial de conhecimento obtido com a coleta, o processamento, o armazenamento e a análise dos dados pode ser utilizado a favor da sociedade.

Por outro lado, os dados podem ser utilizados: pelos governos, para o controle do cidadão e com objetivos políticos; ou pelas organizações privadas, para direcionar um determinado padrão de consumo.

Um caso clássico do uso do *Big Data*, que teve repercussão na mídia em razão da conduta da organização envolvida no episódio, foi a atitude tomada pela rede varejista norte americana *Target* de tentar alterar os hábitos de consumo de suas clientes, por meio de técnicas estatísticas que identificavam a possibilidade de determinada consumidora estar grávida (DAVIS, 2012). Tal fato representa a aplicação de técnicas associadas ao *Big Data*, com implicações éticas e jurídicas.

Azevedo, Neves e Novo (2014) destacam ser necessário que o usuário tenha maior controle sobre quem pode acessar seus dados e o uso que as organizações estão dando a esses dados, no entanto, o tema demanda regulamentações para estabelecer padrões de identificação e de segurança dos dados pessoais.

Um antídoto para condutas antiéticas no ambiente do *Big Data* pode estar disponível no próprio conjunto de normas e regulamentos das organizações, que estabelecem uma série de valores e esclarecem que os colaboradores devem ter confiança e responsabilidade pessoal no processo de análise dos dados (DIAS; VIEIRA, 2013).

As organizações estão experimentando um novo paradigma no qual todos devem levar em consideração questões como a privacidade, a transparência, o rastreamento e como estão sendo utilizados os dados pessoais e corporativos.

Sendo assim, emerge a necessidade crescente de implementar proteções éticas que assegurem a privacidade dos dados.

4. Privacidade

A palavra privacidade, do latim (*privates*), tem o significado de separado do resto, portanto, que uma pessoa pode ficar afastada ou isolada em relação aos demais.

A preocupação com a privacidade antecede a era da internet. O artigo publicado em 1873, pelo juiz americano Tomas Cooley, define a privacidade como a limitação do acesso às informações de uma determinada pessoa, à própria pessoa e à sua intimidade, envolvendo as questões de anonimato, sigilo, afastamento e o direito de ser deixado em paz.

No mundo atual, no qual cada vez mais está presente o uso dos computadores e de mecanismos tecnológicos de comunicação, emerge, segundo Levy (1998), a questão do fim da privacidade e da preservação das informações, decorrente do fluxo informacional produzido e disponibilizado em grande escala na rede mundial de computadores.

A privacidade surge como um desafio no ambiente digital, onde as informações e os dados são gerados, sendo essencial o estudo do tema por parte da Ciência da Informação durante todo o ciclo de vida da informação. A privacidade das informações, de acordo com Smith, Milberg e Burke (1996), é uma das questões éticas da era da informação.

O avanço das tecnologias da informação, os serviços da internet e os *softwares de business intelligence* que realizam a coleta e mineração de grandes quantidades de dados são canais de vulnerabilidade para o acesso às informações (HONG e THONG, 2013).

A era do *Big Data* demanda novos modelos de privacidade. As atividades de identificação das novas informações pessoais são deduzidas por meio de análise preditiva dos dados coletados. Destaca-se a necessidade de inserir a privacidade no contexto do *Big Data*, no qual os indivíduos não só se preocupam com a coleta de dados, mas também com a forma como esses dados serão analisados e usados (MAI, 2016).

Um fator importante de privacidade é a opção de consentimento dada ao consumidor, ou seja, a oportunidade de decidir se o sistema pode ou não usar seus dados. Quando os sistemas de segurança que preservam a privacidade estão funcionando adequadamente, o usuário demonstra confiança para compartilhar as suas informações.

As organizações devem considerar o fato de que a confiança do usuário é mais lucrativa, com resultados positivos a longo prazo, e a quebra dessa confiança terá um impacto negativo. Tratar das preocupações dos usuários em relação à privacidade gera valor para as organizações (MANDIĆ, 2009).

4.1. Disposição para fornecer informações *on-line*

O processo para descobrir padrões de consumo e conhecimento tem tornado a mineração de dados um instrumento de destaque para que as organizações obtenham maior entendimento a respeito dos negócios e do mercado, a partir da análise dos dados minerados no *Big Data*, proporcionando o desenvolvimento de produtos alinhados às necessidades dos usuários. (PROVOST; FAWCETT, 2013).

Na visão do usuário, porém, fornecer informações com base nas suas necessidades e desejos específicos poderá levar a uma possível perda de privacidade (CHELLAPPA; SIN, 2005). A preocupação com a privacidade afeta negativamente a confiança dos usuários nos serviços tecnológicos *on-line* e, consequentemente, a disposição em fornecer suas informações pela internet (MARTINS, 2016).

4.2. As contradições dos usuários

Martins (2016), na discussão dos resultados da sua pesquisa sobre privacidade e confiança, ressalta que, apesar dos usuários acreditarem que fornecer informações pessoais na internet gera riscos, eles encaram que os benefícios trazem compensação. A despeito das profundas preocupações dos usuários com questões de privacidade e segurança, diariamente, os usuários publicam dados nas suas redes sociais.

Segundo Schoenbachler e Gordon (2002), a possibilidade dos usuários fornecerem informações pessoais *on-line* depende do tipo da informação. Os usuários têm mais restrição para informar dados financeiros em comparação com os seus dados demográficos ou de consumo.

A perspectiva de perdas de privacidade e uso indevido de informações no ambiente do *Big Data*, que contempla, por exemplo, o comércio eletrônico, as redes sociais, o *Internet Banking*, os dados do cidadão de posse dos governos, os provedores de internet e as seguradoras, podem influenciar a disposição do usuário em fornecer seus dados (FEATHERMAN; MIYAZAKI; SPROTT, 2010).

Para ter acesso aos serviços *on-line* gratuitos e revolucionários, os usuários concordam em fornecer suas informações sem uma avaliação dos riscos. Os usuários pagam por esses serviços com o que tem de mais precioso: dados pessoais e o seu comportamento no universo *on-line*.

4.3. Riscos à privacidade

A análise dos *Riscos à privacidade* corresponde à avaliação subjetiva: das potenciais perdas de controle sobre a confidencialidade das informações, incluindo as de identificação pessoal;

bem como do uso e da divulgação não autorizados desses dados (FEATHERMAN; MIYAZAKI; SPROTT, 2010).

Nas transações *on-line*, tanto as realizadas no comércio eletrônico como as financeiras, os usuários identificam a falta de informações sobre a privacidade e a potencial perda de controle das informações confidenciais como desvantagens para o uso desses serviços (BELANGER; HILLER; SMITH, 2002).

As organizações que fornecem serviços *on-line* têm a capacidade de coletar dados pessoais confidenciais de alto valor para explorá-los comercialmente (BELANGER; CROSSLER, 2011). Os autores afirmam que ocorrem perdas financeiras e de privacidade dos dados, em razão do uso indevido das informações durante as transações *on-line*.

Entre os fatores que geram vulnerabilidades no ambiente virtual, podem ser destacados os seguintes: i) nas transações *on-line*, os dados do seu computador podem ser comprometidos; ii) a transferência de dados *on-line* pode ser comprometida; iii) transações *on-line* por meio de redes públicas podem trazer riscos; e iv) os dados coletados durante a transação podem ser comprometidos e divulgados sem autorização do usuário (MILNE; CULNAN, 2004).

Para os autores, o risco à privacidade ocorre tanto durante a transação *on-line* como durante o armazenamento das informações do usuário, em razão do fato de as organizações não garantirem que os dados não serão compartilhados ou utilizados no ambiente do *Big Data* para a tomada de decisão.

A falta da privacidade das informações e a sua segurança não estão restritas às empresas que realizam negócios *on-line*. Os dados obtidos pelos governos também estão sujeitos a esses riscos, tanto pela infraestrutura tecnológica desatualizada como pela pouca cultura de Segurança da Informação (SI) nas instituições públicas.

4.4. Preocupação com a privacidade

No ambiente do *Big Data*, a informação trafega com velocidade. Moor (1997) afirma que a informação, quando digitalizada, trafega facilmente e rapidamente no ciberespaço, que é um ambiente resultante da interação de pessoas, *softwares* e serviços da internet, por meio de dispositivos tecnológicos e redes conectadas.

De acordo com o autor, as preocupações com a privacidade emergem quando a velocidade e conveniência fazem com que as informações pessoais tenham uma divulgação não autorizada.

As preocupações dos usuários não ficam restritas somente ao fato de ter uma divulgação não autorizada, mas também ao uso das informações pessoais de forma inadequada e sem permissão.

O avanço das tecnologias da informação, de acordo com Belanger e Crossler (2011), elevou o nível de preocupações com a privacidade das informações, motivando os pesquisadores de sistemas de informação a estudar soluções técnicas para tratar a informação.

A *Internet Privacy Concern* (IPC) é uma área de estudo que, segundo Hong e Thong (2013), tem crescido em decorrência do grande volume de informações que estão sendo coletadas, armazenadas, transmitidas e publicadas na internet, fomentando o ambiente do *Big Data*.

O IPC corresponde ao grau em que o usuário da internet está preocupado com as práticas realizadas pelos *sites* para a obtenção e o uso das informações pessoais (MALHOTRA; KIM; AGARWAL, 2004).

4.5. Confiança no ambiente *on-line*

A confiança se configura como um dos principais fatores que afetam o comportamento dos indivíduos diante de riscos e incertezas. Assim como a privacidade, a confiança é situacional e depende do contexto (LEE TURBAN, 2001).

A forma mais comum de fornecer aos usuários informações para estabelecer uma confiança nos serviços *on-line* diz respeito às declarações e políticas de privacidade. Dessa maneira, as organizações informam aos usuários sobre: serviços disponíveis na internet referentes a sua política de proteção de dados; quem está coletando os dados; e os limites de utilização.

No entanto, os usuários não dispensam seu tempo para ler essas políticas, uma vez que são dispostas em textos longos e escritos com termos jurídicos e técnicos.

Para Mandić (2009), uma forma de aumentar a confiança na privacidade de um serviço *on-line* é usar verificações de selo de privacidade, também conhecido como selo de segurança para site.

O selo de segurança indica que o site tomou medidas de proteção, seja para corrigir vulnerabilidades de segurança ou mesmo para criptografar informações que são trocadas entre o site e os usuários.

5. Segurança da informação

A Associação Brasileira de Normas Técnicas (ABNT), por meio da norma ABNT NBR ISO/IEC 27002:2013, define o termo Segurança da Informação como a proteção da informação de vários tipos de ameaças para garantir a continuidade do negócio, minimizar o risco e maximizar o retorno sobre os investimentos. Definição similar é apresentada por Manoel (2014).

Os princípios básicos da SI - a confidencialidade, a integridade e a disponibilidade - orientam a análise, o planejamento, a implantação e o controle de segurança para as informações das organizações.

As definições desses princípios são: i) *confidencialidade:* proteção das informações contra acesso não autorizado, independente da forma ou do local de armazenamento desses dados; ii) *integridade:* é a proteção de informações, aplicações, sistemas e redes contra mudanças intencionais, não autorizadas ou acidentais; e iii) *disponibilidade:* é a garantia de que as informações e os recursos estão acessíveis aos usuários autorizados, conforme a necessidade (KILLMEYER, 2006).

A gestão da SI envolve as seguintes atividades: i) elaborar uma Política de Segurança da Informação; ii) definir papéis e responsabilidades relacionados com a SI na organização; iii) desenvolver uma estrutura de controle com normas, práticas e procedimentos de SI; iv) estabelecer procedimentos de monitoramento para detectar e assegurar a correção de falhas de segurança; e v) promover a conscientização sobre a necessidade de proteger as informações (WILLIAMS, 2001).

As organizações enfrentam uma revolução nas práticas de gestão da informação, com o foco cada vez maior no valor global das informações protegidas e entregues (*Information Security Governance* – ITGI, 2006).

A segurança, como podemos ver, está relacionada com a capacidade da organização de proteger os dados dos usuários e evitar fraudes *on-line*, por meio de medidas de segurança.

5.1. Riscos à segurança das informações

A preocupação com a gestão adequada da informação no ambiente do *Big Data* envolve o espaço cibernético, que, segundo o autor Killmeyer (2006), é um ambiente propício para a exposição ao risco e no qual estão os ativos de informação, além dos meios de armazenamento, transmissão e processamento dos sistemas de informação.

Segundo Carvalho (2010), o espaço cibernético constitui novo e promissor cenário para a prática de toda a sorte de atos ilícitos, desafia conceitos tradicionais, entre eles o de fronteiras geopolíticas e/ou organizacionais, constituindo novo território, por vezes conhecido e desconhecido, a ser desbravado pelos bandeirantes do século 21.

Os projetos de *Big Data* trabalham com um grande volume de dados provenientes de diversas fontes, que demandam cuidados com a segurança. O armazenamento de um grande volume de dados pode se transformar em alvo de ataques e vazamento de informações sigilosas, o que pode gerar perdas de credibilidade para a organização.

As organizações devem adotar soluções e boas práticas de SI, adequação às normas e leis, definição de políticas, controle de acesso às informações críticas e de capacitação de equipes de TI, entre outras.

Os métodos tradicionais usados para proteger os sistemas de informações contra ameaças de segurança incluem a implementação de *firewalls*, regras de autenticação e o uso de redes privadas virtuais. (AL-SHAWI, 2011). Para o autor, cada uma dessas técnicas tem suas próprias vulnerabilidades e limitações e pode não ser capaz de proteger os recursos de ataques cibernéticos.

Os atacantes coletam e monitoram continuamente os dados dos usuários e das redes governamentais e privadas para tirar vantagem de fraquezas do sistema resultantes de falhas no *design* e na implementação de medidas de segurança, além das falhas ocasionadas em função do baixo nível de maturidade para a organização da informação.

6. Organização da informação

As repetidas ameaças cibernéticas nas organizações de todos os setores, tipos e tamanhos indicam a necessidade da implementação de práticas, métodos e processos relacionados com a organização da informação armazenada.

As instituições estão passando por transformações na forma de lidar com as informações. Considerando o volume de dados disponíveis no *Big Data*, torna-se necessário enfrentar o caos informacional com a utilização da Arquitetura da Informação (AI).

6.1. Arquitetura da informação

A Arquitetura da Informação permite a organização da informação para suporte às ações de gestão do conhecimento, ao mesmo tempo que visa a promover a acessibilidade à informação para a tomada de decisões (LIMA-MARQUES; MACEDO, 2006).

Com base na importância para a organização e apresentação da informação, Richard Saul Wurman utilizou pela primeira vez o termo Arquitetura da Informação em 1976. O criador do termo afirma que o arquiteto da informação dá clareza ao que é complexo, fazendo com que a informação possa ser compreendida (WURMAN, 2005).

Diante de todo esse volume de dados disponível, que pode ser utilizado para a tomada de decisões e melhoria da qualidade de vida, os usuários estão dispostos a trocar informações por serviços melhores, sem a devida atenção sobre as condições de privacidade oferecidas por esses serviços.

Com a frequente evolução de novas ferramentas tecnológicas, a cada dia, o usuário passa a ter mais e mais informação. A informação gerada de forma excessiva, sem critérios de seleção, organização e disseminação, fez surgir, como define Reis (2007), a síndrome da fadiga da informação, caracterizada por tensão, irritabilidade e sentimento de abandono causados pela sobrecarga de informação imposta ao ser humano.

Wurman (1991) afirma que uma edição do *The New York Times* publica, em um dia, mais informações do que um cidadão inglês normal poderia ter recebido durante toda a sua vida, no século 17. O autor adverte que mais dados não significam melhor compreensão, identificando a explosão da não informação.

Toda essa quantidade de informações, para Wurman (1991), leva à síndrome de ansiedade da informação, definida pelo autor como o resultado da distância cada vez maior entre o que compreendemos e o que achamos que deveríamos compreender.

O desenvolvimento e aperfeiçoamento das tecnologias da informação encurtam o caminho do usuário, tanto para obter como para fornecer informações. Todo esse avanço tem as suas vantagens, como também as desvantagens, sobretudo no que se refere à privacidade, à segurança, ao valor e a confiabilidade das informações.

A AI pode ser usada como uma estratégia para a organização da grande massa de informações disponível, para mitigar os riscos relacionados à privacidade, segurança, confiabilidade e perda de valor das informações.

As organizações públicas e privadas têm sido, cada vez mais, cobradas para publicar seus dados brutos em formato eletrônico. No entanto, antes dessa divulgação, visando a mitigar os riscos desse processo, os dados devem ser sanitizados, de modo a haver a remoção de identificadores pessoais. Para isso, podem ser utilizadas técnicas de anonimização (MONTEIRO; MACHADO; BRANCO JR, 2014).

6.2. Anonimização de dados

A anonimização de dados tem um vasto campo de aplicação, podendo ser adotada como medida de segurança. O termo anonimato representa o fato do sujeito não ser unicamente caracterizado dentro de um conjunto de sujeitos. O conceito de sujeito refere-se a uma entidade ativa, como uma pessoa ou computador (MONTEIRO, MACHADO, BRANCO JR, 2014).

O anonimato representa o fato de um registro não ser unicamente identificado em um conjunto de registros. Conjunto de registros pode ser um grupo de pessoas ou rede de computadores (PFITZMANN e KÖHNTOPP, 2005).

Para Camenisch, Fischer-Hübner e Rannenberg (2011), uma transação é considerada anônima quando os seus dados, individuais ou combinados, não possibilitam a associação para identificação de um registro em particular.

Os dados de indivíduos podem ser classificados como:

- Identificadores: atributos que identificam individualmente as pessoas (CPF, nome, identidade);

- Semi-identificadores: atributos que podem ser combinados com informações para reduzir a incerteza sobre a identificação das pessoas (data de nascimento, CEP, profissão, cargo, local de trabalho); e

- Atributos sensíveis: contêm informações sensíveis sobre as pessoas (salário, informações de saúde, despesas de cartão de crédito, hábitos de consumo).

As técnicas que podem ser utilizadas e/ou combinadas para a anonimização dos dados são as seguintes (MONTEIRO; MACHADO; BRANCO JR, 2014):

- Generalização: substitui os valores de atributos semi-identificadores por valores menos específicos e com semântica consistente;

- Supressão: exclui valores de atributos identificadores e/ou semi-identificadores da tabela anonimizada;

- Encriptação: utiliza esquemas criptográficos normalmente baseados em chave pública ou chave simétrica para substituir dados sensíveis por dados encriptados; e

- Perturbação: é utilizada para a substituição de valores dos dados reais por dados fictícios, para mascaramento de banco de dados de testes ou treinamento.

A técnica de perturbação procura alterar randomicamente os dados, com vistas a preservar as características dos dados sensíveis para o modelo de dados, utilizando as seguintes abordagens (CHEN; LIU, 2011):

- Condensação de dados: condensa os dados em múltiplos grupos e tamanhos predefinidos. Dentro de um grupo, não é possível distinguir diferenças entre os registros. Cada grupo tem um tamanho *k*, que é o nível de privacidade decorrente da condensação; e

- *Random Data Pertubation* (RDP): adiciona ruídos, de forma randômica, aos dados sensíveis. A maioria dos métodos utilizados para adicionar ruído randômico corresponde a casos especiais de mascaramento de matriz.

O mascaramento é utilizado na disponibilização de bases de dados para teste ou treinamento, com informações que não identificam os usuários, mas que pareçam ser reais. As técnicas de mascaramento de dados são (LANE, 2012):

- Substituição: substituição randômica de conteúdo por informações sem relação com o dado real;

- Embaralhamento (*Shuffing*): substituição randômica do dado real por um dado derivado da própria coluna da tabela;

- *Blurring*: técnica aplicada a números e datas. Muda o valor do dado por uma percentagem do seu valor original; e

- Anulação/Truncagem: substitui os dados sensíveis por valor nulos (*null*).

6.3. Modelos de anonimização

Diante da necessidade de se manter a privacidade dos dados e a segurança das informações, são apresentados, a seguir, os principais modelos de anonimização encontrados na literatura: *k-anonymity, l-divesity, t-closeness e b-likeness*.

- *k-anonymity*: demanda que qualquer combinação de atributos semi-identificadores seja compartilhada por pelo menos *k* registros, em um banco de dados anonimizado. Este modelo assume o pressuposto de que cada registro representa apenas uma pessoa;

- *l-diversity*: captura o risco da descoberta de atributos sensíveis em um banco de dados anonimizado;

- *t-closeness*: propõe a proteção contra a divulgação de atributo sensíveis; e

- *b-likeness*: apresenta-se como uma solução ao problema, que ocorre com menor frequência, de exposição de privacidade de valores de atributos sensíveis.

Visando a alcançar elevado nível de segurança, podem ser utilizadas ferramentas de segurança para: limitar o acesso aos dados; preservar a privacidade dos usuários; liberar dados úteis para mineradores de dados, sem divulgar as identidades dos usuários; desenvolver modelos de privacidade adequados para quantificar a possível perda de privacidade, em razão de diferentes ataques; e aplicar técnicas de anonimização (LEI XU; WANG; YUAN; REN, 2014).

7. Panorama da proteção de dados em 2018

No ano de 2018, aumentou o foco para a proteção das informações, quando se descobriu que os dados de 87 milhões de usuários do *Facebook* foram utilizados para traçar perfis de comportamento e influenciar politicamente a eleição americana, além do plebiscito que separou o Reino Unido da União Europeia.

Em setembro de 2018, o *Facebook* descobriu um ataque *hacker* que alcançou 50 milhões de usuários em todo o mundo. Em razão desse fato, vários perfis foram desconectados. Uma nova falha, ocorrida em dezembro, possibilitou a exposição das imagens postadas por 6,8 milhões de usuários.

O *The New York Times* revelou, em dezembro do mesmo ano, que o *Facebook* forneceu, sem autorização, dados de usuários a empresas como *Microsoft*, *Netflix*, *Spotify*, *Amazon* e *Yahoo*. As autorizações davam acesso às mensagens privadas. Segundo a reportagem, as empresas podiam ler, escrever e apagar as mensagens, além de ver todos os participantes em um tópico. A reportagem não detalha como isso era feito.

Onde meus dados foram parar? No caso da *Cambridge Analytica*, 300 mil pessoas foram pagas para participar de um teste de personalidade e fornecer seus dados. Elas, porém, foram usadas para coletar dados de outros. Com isso, foi possível criar um banco de dados com 87 milhões de pessoas, que não tinham ideia de que seriam envolvidas em campanhas políticas e outras atividades.

No ambiente do *Big Data*, existe uma fatia considerável de usuários que não se importa em fornecer seus dados nas redes sociais, mas não aceita que suas informações sejam usadas para vender mensagens com as quais não concorda.

Você é o produto: preocupe-se com o que fazem com seus dados.

O *General Data Protection Regulation* (GDPR)[4] regulamenta os direitos dos usuários europeus no que diz respeito à proteção e ao controle de seus dados pessoais. Por meio desse instrumento legal, as pessoas têm direito de saber se seus dados serão usados para gerar propagandas, se as informações serão geradas para criar perfis ou se as empresas que coletam dados vendem ou venderão esses dados a terceiros.

4 Acesse informações sobre o GDPR em https://eugdpr.org/.

No Brasil, a Lei Geral de Proteção de Dados Pessoais (LGPD), de n.º 13.709/2018, estabelece uma série de regras que empresas e outras organizações atuantes no País devem seguir para permitir que o cidadão tenha mais controle sobre o tratamento que é dado às suas informações pessoais.

8. Conclusão

A análise realizada neste artigo permite abordar a problemática acerca dos riscos à privacidade, segurança e organização da informação no ambiente do *Big Data*. Para discutir o tema, este estudo também apresentou questões que possibilitam identificar esses riscos e a preocupação gerada em função deles.

Fica evidente, no contexto do trabalho, o paradoxo dos benefícios que a coleta dos dados apresenta aos usuários, tendo em vista que os avanços tecnológicos estão promovendo: maior facilidade na busca por serviços; e o aumento da exposição dos usuários no espaço cibernético, com risco à privacidade.

No entanto, não é uma tarefa simples, para organizações que utilizam todo o potencial do volume de dados produzidos diariamente pelos usuários no ambiente do *Big Data*, manter um nível de segurança da informação adequado. Organizações de qualquer setor estão sujeitas às ameaças cibernéticas que são disseminadas pelos *hackers*.

Para que todo o potencial do *Big Data* possa ser explorado pelas organizações, é fundamental assegurar a privacidade, segurança e organização das informações. Vários modelos de anonimização que podem ser utilizados para preservar a privacidade dos usuários são propostos na literatura.

O avanço tecnológico não garante uma eficaz segurança da informação, sem uma conscientização do ser humano em relação à segurança. O acesso não autorizado a informações, lugares, objetos, entre outros tipos de dados, na organização, torna a segurança vulnerável, uma vez que as pessoas e as empresas interessadas nesses dados têm acesso indevido a essas informações.

As políticas de privacidade dos serviços *on-line* oferecidos pelas organizações devem estar em conformidade com a LGDP e GDPR. As referidas leis podem aplicar penalidades para as organizações que não se prepararem corretamente para a coleta, a gestão e o uso dos dados privados dos usuários.

Estar *compliance* com a LGDP e GDPR será não só uma oportunidade para melhorar e aumentar o nível de privacidade, segurança e gerenciamento de dados, como um diferencial para os novos modelos de negócio baseados em dados.

Em 2018, emerge o conceito de que somos o produto do espaço cibernético. Grande quantidade de informação é publicada no ciberespaço e os sistemas que recebem esses dados ficam cada vez mais inteligentes, ou seja, são capazes de fazer cruzamentos que nem imaginamos.

O *Big Data* é uma realidade. Os efeitos da tecnologia da informação estão no dia a dia das pessoas, dominando as suas vidas de formas que elas não imaginam.

Referências

AL-SHAWI, A. **Data mining techniques for information security applications.** John Wiley & Sons, Inc., v. 3, May/June 2011.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS – ABNT. **NBR ISO/IEC 27002:2013:** Tecnologia da informação – Técnicas de segurança – Código de prática para a gestão da segurança da informação. Rio de Janeiro: 2013.

AZEVEDO, M.M.; NEVES, J.M.S.; NOVO, R.F. **O crescimento do *Big Data* e as possíveis implicações éticas do seu uso na análise das redes sociais.** *In*: WORKSHOP DE PÓS-GRADUAÇÃO E PESQUISA DO CENTRO PAULA SOUZA, 9., Estratégias Globais e Sistemas Produtivos Brasileiros, 2014.

BELANGER, F.; HILLER, J.S.; SMITH, W.J. Trustworthiness in electronic commerce: the role of privacy, security, and site attributes. **Journal of Strategic Information Systems.** v. 11 n. 3/4, p. 245–70, 2002.

BELANGER, F.; CROSSLER, R.E. Privacy in the Digital Age: A Review of Information Privacy Research in Information Systems. **Mis Quarterly**, v. 35, n. 4, p. 1017–1041, 2011.

BRASIL. Presidência da República. **Lei n.º 13.709, de 14 de agosto de 2018. Lei geral de proteção de dados.** Dispõe sobre a proteção de dados pessoais e altera a Lei n.º 12.965, de 23 de abril de 2014(Marco Civil da Internet). Disponível em: *http://www.planalto.gov.br/ccivil_03/_ Ato2015-2018/2018/Lei/L13709.htm*. Acesso em: 20 dez. 2018.

BRYNJOLFSSON, E; MCAFEE, A. *Big Data* - **a revolução da gestão.** Harvard Business Review, 2012.

CAMENISCH, J.; FISCHER-HÜBNER, S.; RANNENBERG, K. **Privacy and identity management for life.** Springer. 2011.

CARVALHO, P.S.M. **A Defesa cibernética e as infraestruturas críticas nacionais.** Núcleo de Estudos Estratégicos, Comando Militar do Sul, 2010.

CHELLAPPA, R.K.; SIN, R.G. Personalization versus Privacy: an empirical examination of the online consumer's dilemma. **Information Technology and Management**, v. 6, p. 181-202, 2005.

CHEN, K.; LIU, L. Privacy preserving data classification with rotation perturbation. In: IEEE International Conference on Data Mining, 5. IEEE Computer Society, 2011. **Proceedings...** 2011.

CHEN X.S.; YANG L.; LUO Y.G. Large data security protection technology. **Engineering Science and technology,** v. 49, n. 05, p. 1-12, 2017.

COMISSÃO EUROPEIA. **Regulamento Geral sobre a Proteção de Dados.** Disponível em: *https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_pt*. Acesso em: 20 dez. 2018.

DAVENPORT, THOMAS. *Big Data* **at work, uncovering the opportunities**, 2014.

DAVENPORT, THOMAS. **Dados demais! como desenvolver habilidades analíticas para resolver problemas complexos, reduzir riscos e decidir melhor**. 1. ed. Rio de Janeiro: Elsevier, 2014.

DAVIS, K. **Ethics of** *Big Data***.** Sebastopol: O`Reilly Media, 2012.

DIAS, G.A; VIEIRA, A.N. *Big Data*: questões éticas e legais emergentes. **Revista Ciência da Informação,** Brasília, DF, v. 42 n. 2, p.174-184, 2013.

DRINKWATER, D. **Does a data breach really affect your firm's reputation.** Disponível em: *http://www.csoonline.com/article/3019283/data-breach/does-a-data-breach-reallyaffect-your-firm-s-reputation.html*, 2016. Acesso em: 20 dez. 2018.

EREVELLES, S.; FUKAWA, N.; SWAYNE, L. *Big Data* **consumer analytics and the transformation of marketing**. Journal of Business Research, 69(2), 897–904, 2016.

FEATHERMAN, M.S.; MIYAZAKI, A.D.; SPROTT, D.E. Reducing *online* privacy risk to facilitate e-service adoption: the influence of perceived ease of use and corporate credibility. **The Journal of Services Marketing**, v. 24, n. 3, p. 219-229, 2010.

GOLDMAN, A., *et al.* Apache hadoop: conceitos teóricos e práticos, evolução e novas possibilidades. *In*: JORNADAS DE ATUALIZAÇÕES EM INFORMÁTICA, 31., 2012. **Anais...** 2012.

HONG, W.Y.; THONG, J.Y.L. Internet Privacy Concerns: An Integrated Conceptualization and Four Empirical Studies. MIS Quarterly, v. 37, n. 1, p. 275, 2013.

INFORMATION SECURITY GOVERNANCE – ITGI. **Guidance for information security managers**. EUA: 2006.

JANSSEN, M.; VAN DER VOORT, H.; WAHYUDI, A. Factors influencing *Big Data* decision-making quality. **Journal of Business Research**, v. 70, n. 1, p. 338–345, 2017.

JOHNSTON, A.C.; WARKENTIN, M. Fear appeals and information security behaviors: An empirical study. **MIS Quarterly**, v. 34, n. 3, p. 549 – 566, 2010.

KILLMEYER, J. **Information security architecture: an integrated approach to security in organization.** Florida: Auerbach Publications, 2006.

LANE, A. **Understanding and selecting data masking solutions: Creating secure and useful data**, 2012.

LEE, M.K.O.; TURBAN, E. A trust model for consumer Internet shopping. **International Journal of Electronic Commerce**, v. 6, n. 1, p. 75-91, 2001.

LEI XU, C.J.; WANG, J.; YUAN J.; REN, Y. **Information Security in *Big Data*: Privacy and Data Mining.** IEEE, v. 2, 2014

LEVY, P. A **Inteligência Coletiva: por uma antropologia do ciberespaço**. Rio de Janeiro: Loyola, 214 p. 1998.

LIMA-MARQUES, M.; MACEDO, F.L.O. Arquitetura da informação: base para a gestão do conhecimento. In: TARAPANOFF, K. (Org.). **Inteligência, informação e conhecimento em corporações**. IBICT, UNESCO, Brasília, 2006.

MAI, J-E. *Big Data* privacy: The datafication of personal information. **The Information Society**, 2016.

MALHOTRA, N.K.; KIM, S.S.; AGARWAL, J. Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. **Information Systems Research**. v. 15, n. 4, p. 336–355, 2004.

MANDIĆ, M. Privacy and Security in E-Commerce. Art Design and Internet Technologies. **Privatnost I Sigurnost**, v. XXI, br. 2, str. 247–260, 2009.

MANOEL, S.S. **Governança de Segurança da Informação: como criar oportunidades para o seu negócio**. Rio de Janeiro: Brasport, 2014.

MARTINS, R. M. **Preocupação com a privacidade, confiança e disposição dos consumidores a fornecer informações** *on-line* **no contexto do** *Big Data*. Universidade Federal de Uberlândia, 2016.

MILNE, G.R.; CULNAN, M.J. Strategies for reducing online privacy risks: why consumers read (or don't read) online privacy notices. **Journal of Interactive Marketing**, v. 18, n. 3, p. 15-29, 2004.

MONTEIRO, J.M.; BRANCO, E.C. JR; MACHADO, J.C. **Estratégias para Proteção da Privacidade de Dados Armazenados na Nuvem**. Tópicos em Gerenciamento de Dados e Informações, 2014.

MOOR, J. H. Towards a Theory of Privacy in the Information Age. **Computers and Society**, Sep., 1997.

PFITZMANN A; KÖHNTOPP, M. **Anonymity, unobservability, and pseudonymity – a proposal for terminology**. In Designing privacy enhancing technologies, Springer, 2005.

PROVOST, F.; FAWCETT, T. Data science and its relationship to *Big Data* and data-driven decision making. *Big Data*, v. 1, n. 1, p. 51-59, 2013.

REIS, G.A.D. **Centrando a Arquitetura de Informação no usuário.** Escola de Comunicação e Artes, Universidade de São Paulo. São Paulo, 2007.

SCHOENBACHLER, D.D.; GORDON, G.L. Trust and customer willingness to provide information in database-driven relationship marketing. **Journal of Interactive Marketing**, v. 16, n. 3, p. 2-16, 2002.

SHINATAKU, M; DUQUE, C.G.; SUAIDEN, E.J. Análise sobre o uso das tendências tecnológicas nos repositórios brasileiros. **Pesquisa Brasileira em Ciência da Informação e Biblioteconomia**. João Pessoa, v. 9, n. 2, p. 001-012, 2014.

SMITH, H.J.; MILBERG, S.J.; BURKE, S.J. Information Privacy: Measuring Individuals' Concerns About Organizational Practices. **MIS Quarterly**, v. 20, n. 2, p. 167-196, 1996.

WILLIAMS, P.A. Information Security Governance. **Information Security Technical Report**. v. 6, n. 3 p. 60–70, 2001.

WURMAN, R.S. **Information Architects**. Zurich, Schweiz: Gingko Press, 240 p., 1997.

WURMAN, R.S. **Ansiedade de Informação: como transformar informação em compreensão.** São Paulo: Cultura Editores Associados. 1991.

WURMAN, R S. **Ansiedade de Informação 2**. São Paulo: Editora de Cultura, 2005. 298 p. Tradução de Information Anxiety 2, Indianapolis, IN: QUE, 2001. 350 p.

ZWITTER, A. *Big Data* **ethics**. *Big Data* & Society, 2014.