relato de pesquisa

ANÁLISE DO EQUILÍBRIO ENTRE PRIVACIDADE E UTILIDADE NO ACESSO A DADOS

Elaine Parra Affonso[*]
Sandra Cristina de Oliveira[**]
Ricardo César Gonçalves Sant'Ana[***]

RESUMO Este artigo analisa medidas para proteção da privacidade por meio de técnicas não perturbativas e perturbativas, buscando o equilíbrio entre privacidade e utilidade de dados. Assim, utilizou-se de supressão e generalização para anonimização de dados de consultas médicas e, com base em Mivule (2015), realizou-se a adição de ruídos para gerar dados modificados privados. Posteriormente, verificou-se a utilidade destes dados por meio de análise de agrupamento. A generalização mantém a veracidade dos dados e, quando combinada à adição de ruídos, pode ser uma alternativa viável para disponibilizar dados, minimizando ameaças à privacidade e considerando a questão da utilidade.

Palavras Chave: Proteção de dados. Anonimização. Utilidade de dados.

* Doutoranda no Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista – Campus de Marília, Brasil. Professora da Faculdade de Tecnologia de Presidente Prudente, Brasil.
E-mail: elaineaffonso@marilia.unesp.br.

** Doutora em Ciências pela Universidade de São Paulo, Brasil. Professora Assistente Doutor da Universidade Estadual Paulista Júlio de Mesquita Filho - Campus de Tupã, Brasil.
E-mail: sandra@tupa.unesp.br.

*** Doutor em Ciência da Informação pela Universidade Estadual Paulista Júlio de Mesquita Filho, Brasil. Professor Assistente da Universidade Estadual Paulista Júlio de Mesquita Filho, Brasil. Docente do Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista Júlio de Mesquita Filho, Brasil.
E-mail: ricardosantana@marilia.unesp.br.

1 INTRODUÇÃO

A disponibilização de dados por parte das organizações pode contribuir para diversas áreas do conhecimento e para a sociedade, entretanto, a partir desse processo emergem em contraponto, questões tais como à violação da privacidade. Sánchez e Batet (2015) afirmam que devido à natureza confidencial de parte das informações, as organizações responsáveis por elas devem tomar medidas de proteção adequadas para garantir o direito à privacidade do indivíduo. Para Bettini e Riboni (2015) esta violação acontece quando informações pessoais são adquiridas, armazenadas ou processadas por um terceiro sem o consentimento dos indivíduos referenciados.

> A privacidade perfeita pode ser conseguida por meio da publicação de nada, mas isso não tem utilidade; a utilidade perfeita pode ser obtida por meio da publicação de dados exatamente como se recebeu, mas esta não oferece privacidade (DWORK, 2008, p.6, tradução nossa).

A busca é pelo equilíbrio entre privacidade e utilidade dos dados ao disponibilizá-los, visto que a proteção da privacidade tornou-se uma questão importante na relação entre a sociedade e aqueles que detêm os elementos chave no fluxo informacional. Para Wong et al. (2006) e Run et al. (2012) cada vez mais as organizações têm que disponibilizar dados para a sociedade e quando os dados podem ser combinados com outras bases de dados públicas é necessário utilizar algum método para garantir a privacidade do sujeito.

Mivule (2015) também corrobora que as organizações têm interesse em colaborar com pesquisas que envolvem compartilhamento de dados, todavia, muitas vezes esta ação pode trazer riscos de confidencialidade. Assim, muitas organizações compartilham dados com os identificadores pessoais suprimidos e vários métodos têm sido propostos para proteção da privacidade, contudo, a utilidade do conjunto de dados privados continua sendo um desafio, pois ao utilizar métodos de anonimização pode-se também diminuir a utilidade dos dados.

A utilidade dos dados *versus* a privacidade está diretamente ligada em quão útil um conjunto de dados disponibilizado é para o usuário, de forma a satisfazer sua necessidade de consulta (MIVULE, 2013). Esta satisfação tem relação com o significado que esses dados irão trazer para o usuário, tornando possível a apropriação da informação sobre determinado contexto, pois não basta estar disponível, precisa ser útil.

Nesse contexto, a Ciência da Informação tem muito a contribuir com estudos que permitam atender a demanda pela busca e disponibilização dos dados, a fim de gerar informação para o usuário, pois a Ciência da Informação se preocupa com "[...] a análise dos processos de construção, comunicação e uso da informação" (LE COADIC, 1996, p. 26) e "[...] relaciona com o corpo de conhecimentos relativos à produção, coleta, organização, armazenamento, recuperação, interpretação, transmissão, transformação e utilização da informação" (BORKO, 1968, p. 3, tradução nossa).

Não é relevante disponibilizar um conjunto de dados se estes não apresentarem utilidade para o usuário, pois "o objetivo final de um produto de informação, de um sistema de informação, deve ser pensado em termos de usos dos dados à informação e dos efeitos resultantes desses usos nas atividades dos usuários" (LE COADIC, 1996, p. 39).

Em um processo de divulgação dos dados, o detentor de dados é o responsável pelo conjunto de dados originais. Esses dados estão no estado bruto e apresentam utilidade integral, pois não passaram por nenhum processo de tratamento dos dados, entretanto, não existe o conceito de privacidade. Para garantir a privacidade do sujeito referenciado nesses dados, podem ser utilizadas técnicas de anonimização, tais como: supressão, generalização, adição de ruídos ou troca de dados (*swapping*). Por meio destas técnicas, obtém-se um conjunto de dados anonimizados, pois ao divulgar o conjunto de dados é possível ter acesso aos dados do sujeito, contudo, a sua identidade e privacidade mantêm-se protegidas.

Assim, no processo de proteção da privacidade, observa-se de um lado o detentor de dados e de outro o conjunto de dados anonimizados, que é permeado pela díade utilidade dos dados e privacidade dos sujeitos referenciados por estes dados, visto que o processo busca o equilíbrio entre estas duas forças antagônicas. Logo, quanto maior a proteção dos dados, maior tendência à privacidade e à probabilidade de redução da utilidade dos dados disponibilizados. Uma proteção dos dados mínima pode garantir e manter a utilidade dos dados disponibilizados, no entanto, pode significar riscos ou violação da privacidade dos indivíduos envolvidos.

> Assim, a iniciativa de disponibilizar e compartilhar dados emerge preocupações de privacidade, pois, os dados devem ser anonimizados antes de serem publicados, que deve não apenas satisfazer as exigências de privacidade, mas também a utilidade dos dados (NERGIZ; GÖK, 2014, p. 1. tradução nossa).

Nesse sentido, aumentam-se as indagações sobre a possibilidade de proteger dados pessoais ao permitir a disponibilização e ao mesmo tempo manter a utilidade. Adams e Sasse (2001) ainda relatam que muitas das violações de privacidade são devido à incapacidade de antecipar como esses dados poderiam ser utilizados, por quem e como isso afetaria os usuários.

O objetivo deste trabalho foi analisar medidas para proteção da privacidade por meio de técnicas não perturbativas e perturbativas, a fim de gerar um conjunto de dados anonimizados como alternativa para minimizar quebras de privacidade e manter a utilidade destes.

Busca-se com este estudo, mostrar que, ao utilizar conjuntamente técnicas não perturbativas e perturbativas para proteção da privacidade, pode ser possível disponibilizar o acesso e uso de dados, minimizando questões vinculadas à privacidade e ainda, com níveis aceitáveis de utilidade. Ressalta-se, no entanto, que não é alvo deste trabalho, verificar o grau de anonimização e realizar ataques no conjunto de dados anonimizados.

Este texto é apresentado da seguinte forma: investigação bibliográfica, que se baseou

nas técnicas de proteção da privacidade; realização do processo de anonimização de dados de consultas médicas por meio da aplicação de técnicas não perturbativas de generalização e supressão nos dados de variáveis qualitativas; utilização de técnicas perturbativas propostas por Mivule (2015), ou seja, de adição de ruídos e uso de transformada da distância nos dados de variável quantitativa, resultando em um conjunto de dados modificados privados; e, verificação da utilidade por meio de método de agrupamento, seguindo o modelo de Mivule (2015).

2 PROTEÇÃO DA PRIVACIDADE E ANONIMIZAÇÃO DE DADOS

Tendo em vista a necessidade de compreender os aspectos envolvidos na proteção de dados pessoais, faz pertinente a discussão de conceitos e técnicas no âmbito da privacidade.

Para Westin (1967, p.5, tradução nossa) a privacidade pode ser definida como "[...] o direito de indivíduos, grupos e instituições determinarem quando, como e quais informações sobre eles serão divulgadas a outros" e, acrescenta que a privacidade pode compreender diferentes estados, tais como: solidão, intimidade, reserva e anonimato.

O anonimato é uma condição na qual algo relacionado a um indivíduo pode ser conhecido, mas não pode ser vinculado à identidade de um indivíduo específico. Desta forma, ao buscar a privacidade, a identidade do sujeito pode ser revelada, mas não se tem conhecimento das informações associadas a ele, entretanto, no anonimato, tem-se o conhecimento das informações relacionadas a um sujeito, mas não é divulgada sua identidade (SKOPEK, 2014).

Com o objetivo de minimizar ameaças à privacidade dos sujeitos, os detentores de dados têm realizado a remoção de dados identificadores únicos, tais como nome e número de documentos pessoais, e assim, acreditam que o anonimato é garantido, pois os dados resultantes após esta remoção geram um olhar anônimo. Entretanto, este conjunto de dados teoricamente anonimizados pode ser combinado com outras bases de dados, ameaçando a privacidade do indivíduo (RUN *et al*., 2012; SAMARATI; SWEENEY, 1998).

Logo, ampliam-se a necessidade de buscar alternativas de proteger a privacidade do indivíduo, cuja finalidade, segundo Sánchez e Batet (2015), é diminuir a probabilidade de identificação de um sujeito em um conjunto de dados ou conhecer seus atributos confidenciais e, ainda para Ciriani *et al.* (2010), uma questão relevante é a proteção da identidade do usuário vinculada ao conjunto de dados, mantendo seu anonimato, pois o anonimato não significa que nenhuma informação foi liberada, mas que as informações divulgadas não podem ser relacionadas a um único sujeito.

As principais técnicas para anonimizar dados e garantir a proteção da privacidade incluem: generalização, supressão, permutação e perturbação de dados (FUNG *et al*., 2010).

Para De Capitani di Vimercati *et al.* (2012) as definições de privacidade, como também as técnicas para proteção de dados podem ser classificadas em duas categorias: privacidade sintática e privacidade semântica.

Na privacidade sintática o conjunto de dados normalmente é disponibilizado na forma de tabela e, são caracterizados em: identificadores (identificam unicamente o indivíduo); semi-identificadores (atributos que, quando combinados com outras fontes de dados externas podem identificar o indivíduo); sensíveis (representam dados confidenciais do sujeito); e, não confidenciais (não apresentam valores confidenciais e cuja divulgação é inofensiva). Ao utilizar esta categoria, o primeiro passo consiste em remover os atributos identificadores, todavia, observou-se que apenas a supressão não é o suficiente para garantir o anonimato de um conjunto de dados, pois os atributos semi-identificadores podem ser correlacionados com fontes externas de dados e revelar informações confidenciais sobre o sujeito (DE CAPITANI DI VIMERCATI *et al*., 2012).

Observa-se em De Capitani Di Vimercati *et al*. (2012) que na categoria privacidade sintática prevalece à utilização de métodos não-perturbativos, o qual para Ciriani *et al.* (2007), objetiva-se em não modificar os valores dos dados originais durante o processo de proteção da privacidade, embora elimine detalhes dos valores originais. Incluem nesta classificação modelos para proteção da privacidade, tais como: k-anonimato, *l-diversity*, t-*closeness* e, técnicas como a supressão e generalização dos dados (DE CAPITANI DI VIMERCATI *et al.,* 2012; MIVULE, 2014).

A generalização, uma técnica não-perturbativa (CIRIANI *et al*, 2007), consiste em substituir os valores dos atributos semi-identificadores por valores menos específicos, mas mantendo a

representação semântica desses dados, enquanto que a supressão "tem a finalidade de remover da tabela de dados, uma célula, uma coluna, uma tupla ou um conjunto dos mesmos" (CIRIANI *et al.*, 2010, p. 10, tradução nossa).

Para Nergiz e Gök (2014) uma das vantagens da generalização é preservar a veracidade dos dados e, De Capitani di Vimercati *et al.* (2012) relatam que a combinação de generalização e de supressão pode disponibilizar informações mais verdadeiras, embora, não completas, isso devido à redução dos detalhes dos dados que serão divulgados.

Com base nessas reflexões adota-se neste estudo **privacidade sintática** como conjunto de técnicas que tem como foco aspectos estruturais e de composição dos dados, visando impacto mínimo na semântica do conteúdo dos atributos, na busca por uma redução do potencial de identificação dos envolvidos no conjunto de dados alvo.

A privacidade semântica utiliza técnicas que visam garantir que uma tabela de dados não divulgue informações sensíveis sobre o indivíduo (DE CAPITANI DI VIMERCATI *et al.*, 2012). Assim, na privacidade semântica dos dados observa-se a utilização de técnicas perturbativas, que para Mivule (2014) consiste em transformar ou perturbar valores dos dados originais por meio de adição de ruído, ruído multiplicativo, ruído multiplicativo logarítmico ou privacidade diferencial.

Portanto, adota-se o conceito de **privacidade semântica** como o conjunto de técnicas utilizadas na modificação do conteúdo dos atributos (significado), buscando a proteção de dados confidenciais e, por meio destas técnicas pode-se gerar um conjunto de dados modificados.

Por meio desta classificação, De Capitani Di Vimercati *et al.* (2012, p.5, tradução nossa) concluem que:

> As técnicas sintáticas para proteção dos dados normalmente propendem a manter a veracidade das informações divulgadas, enquanto que as técnicas semânticas normalmente adicionam ruídos, perturbando o conjunto de dados original e, alcançando assim a privacidade, ao preço da veracidade.

Assim, a finalidade das técnicas para proteção da privacidade é essencialmente evitar que um atacante correlacione dados sensíveis à identidade de um sujeito em um determinado contexto.

3 ADIÇÃO DE RUÍDOS E TRANSFORMADAS DA DISTÂNCIA APRESENTADAS EM MIVULE (2015)

Mivule (2015) utiliza o método de perturbação com adição de ruídos para proteção da privacidade. Este método consiste em substituir valores de um conjunto de dados reais por dados fictícios (valor aleatório – ruído), assim, antes do detentor publicar os dados, ele os altera aleatoriamente, de modo a disfarçar os dados sensíveis, preservando as características estatísticas dos dados (DOMINGO-FERRER; SEBÉ; CASTELLÀ-ROCA, 2004; KARGUPTA *et al.*, 2005). A adição de ruídos é representada por meio da expressão (1) (KIM, 1986).

$$Z = X + \varepsilon \qquad (1)$$

onde: X representa o conjunto de dados numéricos originais e ε é o conjunto de valores aleatórios (ruídos), obtido a partir de uma distribuição Gaussiana $\varepsilon \sim N(\mu, \sigma^2)$, ou seja, com média (μ) e desvio padrão (σ^2). Os valores desta distribuição são adicionados a X, e, finalmente, o símbolo Z representa o conjunto de dados modificados (KARGUPTA *et al.*, 2005; MIVULE, 2013).

Mivule (2015) apresenta ainda um conjunto de dados modificados privados, empregando uma combinação de medidas para proteção da privacidade e técnica de transformada da distância, por meio da distância euclidiana, buscando manter algumas características estatísticas dos dados originais. Para alcançar melhores resultados, o autor utiliza filtro de média móvel e, posteriormente é realizada análise da utilidade dos dados por meio do método de agrupamento k-média e índice de *Davies-Bouldin* para avaliar o desempenho do agrupamento (*cluster*).

A análise de agrupamento por meio do algoritmo k-média consiste no método de agrupamento não hierárquico, onde k representa a quantidade de centroides (centro de cada agrupamento) que irá agrupar por similaridade. Segundo Corrar, Paulo e Dias Filho (2009, p. 348), o k-média procura:

> diretamente uma partição de n objetos, de modo que satisfaça à duas premissas básicas: coesão interna e isolamento

dos grupos. O processo consiste em primeiramente selecionar um grupo de origem (ou semente) como o grupo central inicial, e todos os objetos (indivíduos), dentro de uma distância inicial pré-estabelecida, são incluídos nos grupos resultantes. Então, outro grupo origem é escolhido e a designação continua até que todos os objetos sejam distribuídos.

4 PROCEDIMENTOS METODOLÓGICOS

Para realizar o processo de anonimização e proteção da privacidade, foram consideradas as seguintes definições:

a) Privacidade sintática, por meio de técnicas não perturbativas de supressão e generalização;

b) Privacidade semântica, por meio de técnica perturbativa baseada em Mivule (2015), que considerou especificamente a adição aleatória de ruídos e cálculo da transformada da distância para geração de um conjunto de dados modificados privados, incluindo a verificação da utilidade dos dados por meio de análise de agrupamento.

Para o desenvolvimento deste trabalho, utilizou-se como exemplo um conjunto de dados com valores simulados de sujeitos envolvidos em consultas médicas. A base de dados foi estruturada a partir dos elementos da guia de consulta definida no padrão Troca de Informação da Saúde Suplementar – TISS (ANS, 2016), disponibilizada pela Agência Nacional de Saúde – ANS (Quadro 1). O acesso aos elementos da guia de consulta no sítio da ANS ocorreu no mês de junho de 2016.

Quadro 1 - Estrutura de atributos de guias de consultas médicas

Atributos	Descrição	Preenchimento
RegANS	Registro da operadora de plano privado de assistência à saúde na ANS.	Obrigatório
NGuiaPrest	Nº que identifica a guia no prestador de serviços.	Obrigatório
NGuiaOp	Nº que identifica a guia atribuída pela operadora.	Condicionado
ValCart	Validade da carteira.	Obrigatório
NumCart	Nº da carteira do beneficiário na operadora.	Obrigatório
RN	Indica se o paciente é um recém-nato que está sendo atendido no contrato do responsável.	Obrigatório
Benef	Nome do beneficiário.	Obrigatório
CNS	Nº do Cartão Nacional de Saúde do beneficiário.	Condicionado
CodOp	Código identificador do prestador contratado executante junto à operadora, conforme contrato estabelecido.	Obrigatório
NomeContr	Razão Social, nome fantasia ou nome do prestador contratado da operadora que executou o procedimento.	Obrigatório
CNES	Código do prestador executante no Cadastro Nacional de Estabelecimento de Saúde (CNES) do Ministério da Saúde.	Obrigatório
NomeProf	Nome do profissional que executou o procedimento.	Condicionado
CProf	Código do conselho profissional do executante do procedimento, conforme tabela de domínio nº 26*.	Obrigatório
NumCons	Nº de registro do profissional executante no respectivo conselho profissional.	Obrigatório
UF	Sigla da Unidade Federativa do Conselho Profissional do Executante do procedimento, conforme tabela de domínio nº 59.	Obrigatório
CBO	Código na Classificação Brasileira de Ocupações do profissional executante do procedimento, conforme tabela de domínio nº 24.	Obrigatório
IndAc	Indica se o atendimento foi devido o acidente ocorrido com o beneficiário ou doença relacionada, conforme tabela de domínio nº 36.	Obrigatório
DtAtend	Data em que o atendimento/procedimento foi realizado.	Obrigatório
Con	Código do tipo de consulta realizada, conforme tabela de domínio nº 52.	Obrigatório
Tab	Código da tabela utilizada para identificar os procedimentos realizados ou itens assistenciais utilizados, conforme tabela de domínio nº 87.	Obrigatório
CodProc	Código identificador do procedimento realizado pelo prestador, conforme tabela de domínio.	Obrigatório
VlProc	Valor unitário do procedimento realizado.	Obrigatório
OBS	Campo utilizado para adicionar quaisquer observações sobre o atendimento ou justificativas que julgue necessário.	Opcional

Fonte: Adaptado da ANS (2016)

* As tabelas de domínio encontram-se no item Componente de Representação de Conceitos em Saúde. Disponível em: http://www.ans.gov.br/prestadores/tiss-troca-de-informacao-de-saude-suplementar/padrao-tiss-setembro-2016

Para alcançar os objetivos propostos, foram efetivadas as seguintes etapas:

a) Supressão dos atributos identificadores que são considerados únicos na tabela e generalização dos semi-identificadores que possuem valores do tipo data;
b) No atributo valor do procedimento (VlProc) foram adicionados valores aleatórios (ruídos) por meio da distribuição normal gaussiana;
c) A partir dos dados com ruídos foi aplicada a transformada da distância, considerando a distância euclidiana (distância unidimensional) para extrair os coeficientes, sendo que estes são adicionados ao conjunto de dados ruidosos e assim, é gerado o conjunto de dados modificados privados;
d) Análise estatística descritiva (elaboração de gráficos e tabelas, e cálculo de medidas descritivas) de ambos os conjuntos de dados, originais e privados;
e) Cálculo da covariância e da correlação linear de Pearson[1] para verificar a similaridade entre os conjuntos de dados, originais e privados;
f) Para reduzir o excesso de ruídos e, alcançar melhores resultados, foi utilizado o filtro de média móvel[2] no conjunto de dados modificados privados;
g) Posteriormente, realizou-se agrupamento dos dados de cada conjunto e, por meio do método de *Ward* [3], foi identificado o número ideal de *clusters* a serem considerados para o agrupamento;
h) Finalmente, aplicou-se o algoritmo k-média para realizar o agrupamento dos dados, como também o cálculo da distância dos elementos ao centroide[4]; e por meio do índice de *Davies-Bouldin*[5] foi verificado o desempenho do agrupamento.

Para a demonstração das técnicas de anonimização, o estudo se baseou na anonimização de dados no contexto do paciente e, embora seja previsto a generalização de todos os atributos semi-identificadores nas questões de anonimização, neste estudo a generalização foi aplicada especificamente nos semi-identificadores Data de Atendimento (DtAtend) e Validade da Carteira (ValCart). O volume de dados simulados foi o suficiente para representar os aspectos relacionados às técnicas utilizadas.

5 RESULTADOS E DISCUSSÕES

Estruturou-se uma base de dados com uma amostra de 400 valores simulados representando os dados coletados de guias de consultas médicas, nos quais foram aplicadas técnicas não perturbativas e perturbativas.

A possibilidade da disponibilização desses dados é contribuir com pesquisadores e oferecer informações tais como tipo de consulta, ocorrência de acidentes, procedimentos realizados e valores gastos nos procedimentos. Contudo, ao permitir a disponibilização e acesso aos dados, tornando os úteis para sociedade, surgem questões que podem comprometer a privacidade do indivíduo, mesmo que os dados identificadores sejam removidos do conjunto de dados.

A primeira etapa para anomização dos dados é realizar as atividades de supressão dos identificadores únicos e generalização de semi-identificadores, de acordo com a proposta idealizada por Samarati e Sweeney (1998).

5.1 Aplicação da supressão e generalização

O Quadro 2 apresenta resultado das operações de supressão e generalização, que são técnicas não perturbativas utilizadas como meio para proteger a privacidade do sujeito, com a finalidade de anonimizar o conjunto de dados.

1 A covariância é uma medida não padronizada da relação linear existente entre dois conjuntos de dados de variáveis quantitativas. Se ela tiver um sinal positivo, indica que as variáveis movem juntas, e um sinal negativo, que elas movem em direções opostas. De forma complementar, a correlação é uma medida padronizada desta associação linear, variando entre -1 e 1. Quanto mais próxima de 1 ou de -1 estiver, maior será a correlação linear positiva ou negativa, respectivamente, entre as variáveis. Quanto mais próxima de zero estiver, menor será a correlação linear entre estas (MARTINS, 2002).

2 O filtro de média móvel é usado para eliminar ou diminuir algum ruído indesejável em um conjunto de dados. É obtido calculando-se a média destes dados, sempre adicionando um novo valor ao conjunto e descartando o mais velho (MORETTIN; TOLOI, 2004).

3 Agrupamento hierárquico que se "baseia na perda de informação decorrente do agrupamento de objetos em conglomerados [...]". "[...] esse procedimento tende a combinar grupos com um menor grupo de observações" (CORRAR; PAULO; DIAS FILHO, 2009).

4 Valores médios das observações sobre as variáveis.

5 Métrica desenvolvida por Davies e Bouldin (1979) para avaliar um agrupamento de dados. Quanto menor for o valor deste índice, melhor será o desempenho do agrupamento.

Foram suprimidos os atributos identificadores (que identificam o indivíduo), ou seja: Nº que identifica a guia no prestador de serviços (NGuiaPrest); Número da guia atribuído pela operadora (NGuiaOp); Número da carteira (NumCart); Nome do beneficiário (Benef); Número do CNS (CNS). Neste trabalho considerou-se a técnica de supressão por coluna.

Posteriormente, foi aplicada a generalização dos atributos semi-identificadores (atributos que quando combinados com outras bases de dados podem possibilitar a identificação do indivíduo) que possuem valores do tipo data, para demonstrar o uso de técnica não perturbativa (Quadro 2).

Quadro 2 - Recorte do conjunto de dados com operações de Supressão e Generalização dos atributos

RegANS	ValCart	RN	CodOp	NomeContr	CNES	NomeProf	Cprof	NumCons	UF	CBO	IndAc	DtAtend
373357	out/18	sim	1000	Hosp. X	2080516	AAA	6	10723	SP	225250	1	jun/15
227326	fev/16	não	2000	Hosp. Y	2053462	BBB	6	17634	SP	225121	9	jan/15
373357	fev/16	não	1000	Hosp. X	2080516	CCC	6	15432	SP	225175	9	jun/15
227326	out/18	não	2000	Hosp. Y	2053462	DDD	6	19573	SP	578934	9	jan/15

Fonte: Elaborado pelos autores

O Quadro 2 apresenta os dados da guia de consulta sem os atributos identificadores e com os atributos com valores do tipo data generalizados. Por exemplo, o atributo data de atendimento (Dt_Atend) que apresentava o valor 13/jun/15, após a generalização é representado pelo valor jun/15. Assim, a generalização consiste em suprimir o valor do dia na data de atendimento. A vantagem da generalização é manter a representação semântica dos valores no conjunto de dados e apresentar um maior número de registros com valores similares, assim dificultando a identificação do sujeito.

5.2 Adição aleatória de ruído

No conjunto de dados com os valores do procedimento (VlProc) foi aplicado o método perturbativo por meio da adição aleatória de ruídos e cálculo da transformada da distância, devido este atributo ser suscetível a quebras de confidencialidade.

Assim, foram gerados valores aleatórios (ruídos) por meio da distribuição Gaussiana, ou seja, $\varepsilon \sim N(\mu, \sigma^2)$, onde foram considerados média (μ) = R$ 68,17 e desvio padrão (σ) = R$ 14,42, que são valores proporcionais à média e ao desvio padrão considerados em Mivule (2015). Após a geração dos ruídos, estes foram acrescentados ao conjunto de dados originais.

Posteriormente, utilizou-se da transformada da distância, por meio da métrica distância euclidiana unidimensional, para extrair os coeficientes dos conjuntos de dados e adicioná-los ao conjunto de dados ruidosos, assim, gerando o conjunto de dados modificados privados.

Ao realizar a estatística descritiva dos conjuntos de dados, observa-se uma dissimilaridade de R$ 67,85 no valor da média dos dados originais em relação aos dados com ruídos, e uma dissimilaridade de R$ 135,71 dos dados originais quando comparados aos dados modificados privados (Tabela 1). A dissimilaridade entre os dados modificados privados e os dados originais pode indicar uma garantia de privacidade.

Tabela 1 - Medidas descritivas dos conjuntos de dados

Dados Originais		Dados com Ruídos		Dados Modificados Privados	
Média	R$ 398,40	Média	R$ 466,25	Média	R$ 534,11
Mediana	R$ 398,00	Mediana	R$ 465,96	Mediana	R$ 532,98
Desvio padrão	R$ 59,70	Desvio padrão	R$ 61,57	Desvio padrão	R$ 66,53
Coef. de variação	0,1498	Coef. de variação	0,1321	Coef. de variação	0,1246
Mínimo	R$ 301,00	Mínimo	R$ 345,25	Mínimo	R$ 388,50
Máximo	R$ 499,00	Máximo	R$ 605,87	Máximo	R$ 712,74

Fonte: Elaborado pelos autores

Entretanto, ao comparar os coeficientes de variação[6] dos três conjuntos de dados, pode-se observar que estes mantêm praticamente a mesma dispersão (baixa dispersão) ao redor de suas respectivas médias, dando indícios de similaridade entre os dados originais, os dados com ruídos e os dados modificados privados.

Além disso, ao realizar o cálculo da correlação linear de Pearson (Tabela 2) entre os dados originais e os dados com ruídos, estes apresentam uma correlação de 0,97, enquanto a correlação dos dados originais com os dados modificados privados é de 0,90. A correlação linear de Pearson indica a similaridade entre as duas variáveis e, como este valor é positivo, as duas variáveis tendem a se mover juntas linearmente. Além disso, a relação entre estas é forte, pois a correlação se aproxima de 1, em ambos os casos (Tabela 2).

A covariância entre os dados originais e os dados com ruídos apresenta o valor de 3.566,84, enquanto os dados originais e os dados modificados privados apresentam covariância de 3.578,47 (Tabela 2). Como o valor da covariância é positivo, confirma-se que os dois conjuntos de dados se movem na mesma direção, em ambos os casos (vide Tabela 2 e Figura 1).

Tabela 2 - Métricas entre dados originais e privados

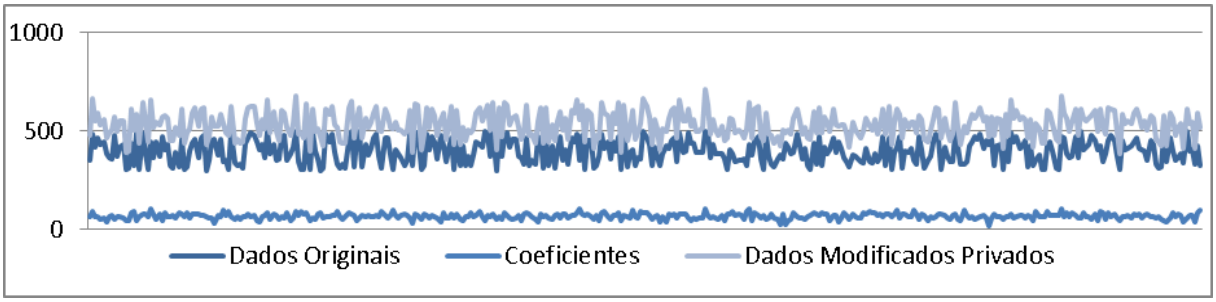
Correlação dados originais versus dados com ruídos	0,97
Correlação dados originais versus dados modificados privados	0,90
Covariância dados originais versus dados com ruídos	3.566,84
Covariância dados originais versus dados modificados privados	3.578,47

Fonte: Elaborado pelos autores

A correlação e a covariância foram utilizadas como medidas estatísticas para verificar se os conjuntos de dados se movem juntos, pois como cita Le Coadic (1996, p.53) "não há ciência ou tecnologia sem medidas, e principalmente sem medidas exatas". No contexto deste trabalho, estas métricas contribuíram para afirmar que o conjunto de dados modificados privados herdou as mesmas características estatísticas dos dados originais, indicando maior utilidade dos dados. Observa-se na Figura 1 que o conjunto de dados modificados privados tende a se mover na mesma direção do conjunto de dados originais e, para Mivule (2015), por manter as características estatísticas, os dados modificados privados apresentam um bom nível de utilidade.

6 Coeficiente de variação é uma medida descritiva obtida a partir da razão entre o desvio padrão e a média de um conjunto de dados, indicando quão dispersos estão os dados em relação à média. Este coeficiente varia entre zero e um e, quanto menor ele for, menor será a dispersão dos dados ao redor da média (MARTINS, 2002).

Figura 1 - Comparação entre os dados originais e os dados modificados privados



Fonte: Elaborado pelos autores

Portanto, ao gerar um conjunto de dados modificados privados para o atributo (Vl_Proc), é possível preservar a mesma estrutura estatística dos dados originais.

5.3 Verificação da utilidade dos dados

Para verificar a utilidade do conjunto de dados modificados privados foi aplicado o método de aprendizagem não supervisionada por meio do algoritmo de agrupamento k-média. Para este procedimento, considerou-se um número de agrupamentos (*clusters*) igual a 3 (k = 3), o qual foi definido com base no dendograma resultante da aplicação do método de *Ward* aos conjuntos de dados em estudo (originais e modificados privados) .

A Tabela 3 ilustra as medidas de desempenho dos agrupamentos por meio da média da distância dos elementos ao centroide, sendo possível avaliar o desempenho do agrupamento (quanto menor for esta média, melhor será o desempenho deste agrupamento). Logo, ao gerar os dados modificados privados observou-se que a média geral apresentou valor maior que a obtida para os dados originais. Assim, para otimizar este resultado, foi aplicado ainda o filtro de média móvel (de ordem dois) no conjunto de dados modificados privados. Consequentemente, a média geral da distância ao centroide para o conjunto de dados filtrados resultou em um menor valor que os demais e, desta forma, pode-se concluir que o conjunto de dados modificados privados filtrados apresentou um melhor desempenho no agrupamento.

Para corroborar a afirmação citada, foi aplicado o índice de *Davies-Bouldin*, onde se obteve para os dados originais o valor de 0,568, para os dados modificados privados 0,975 e para os dados modificados privados filtrados 0,539. Ao considerar que quanto menor o valor do índice de *Davies-Bouldin* melhor será o agrupamento, então, pode-se confirmar que o conjunto de dados modificados privados filtrados teve um melhor agrupamento.

Tabela 3 - Medidas de avaliação do agrupamento

Avaliação da distância do *cluster*	Dados originais	Dados modificados privados	Dados modificados privados filtrados
Média da distância ao centroide 1	17,14	23,09	17,04
Média da distância ao centroide 2	16,18	18,97	17,12
Média da distância ao centroide 3	16,69	20,76	13,43
Média geral da distância ao centroide	16,67	21,00	15,43
Índice de *Davies-Bouldin*	0,568	0,975	0,539

Fonte: Elaborado pelos autores

Em Mivule (2015) o método de agrupamento k-média foi utilizado para verificar a quantidade de elementos em cada conjunto de dados, pois, segundo o autor, para indicar um bom nível de utilidade dos dados, espera-se que a quantidade de elementos em cada *cluster* seja similar em cada conjunto de dados.

Neste trabalho foram obtidas quantidades de elementos próximas nos três *clusters*, tanto para o conjunto de dados originais como para os conjuntos de dados modificados privados e filtrados, indicando um bom nível de utilidade destes dados, e por apresentar esta pequena diferença no número de elementos, é possível garantir ainda uma maior privacidade dos dados, pois o atacante tem mais dificuldades em descobrir o número de elementos em cada *cluster* (Tabela 4). Desta forma, os resultados alcançados neste trabalho corroboram com a análise realizada por Mivule (2015).

Tabela 4 - Número de elementos em cada *cluster*

	Dados originais	Dados modificados privados	Dados modificados privados filtrados
Cluster 1	140	131	111
Cluster 2	136	117	109
Cluster 3	124	152	180
Total	400	400	400

Fonte: Elaborado pelos autores

Ao aplicar os métodos de proteção da privacidade observa-se que o uso simultâneo das técnicas de generalização e de supressão, propostas neste trabalho, associadas ao procedimento de ruídos com as transformadas da distância , proposto por Mivule (2015), pode ser uma alternativa viável para minimizar as ameaças à privacidade e manter a utilidade dos dados quando da divulgação destes.

6 CONSIDERAÇÕES FINAIS

Atender a demanda pela busca de dados e utilizar medidas adequadas para proteção destes é uma questão relevante para detentores de dados, uma vez que o ato de suprimir dados que revelam informações pessoais não garante privacidade ao sujeito e, o ato de aplicar supressão demasiadamente implica na perda da utilidade dos dados. Assim, conseguir o equilíbrio entre privacidade e utilidade de dados reflete na garantia de disponibilizar dados de forma que minimize ameaças à privacidade do sujeito e, simultaneamente, apresentem valor para o usuário.

Assim, quando a questão é a divulgação de dados pelas organizações, observa-se uma busca pelo equilíbrio entre a privacidade e a utilidade de dados, visto que as organizações têm a necessidade de "divulgar" dados e a intenção de proporcionar dados "úteis", ainda com garantia de "privacidade" aos sujeitos referenciados em um conjunto de dados.

Entre estes elementos emerge o processo de proteção da privacidade, cujas técnicas para esta finalidade têm mostrado que, quando se atinge a privacidade, a tendência é perder a utilidade e, ao alcançar a utilidade, aumenta os aspectos de violação da privacidade. Assim, pesquisadores buscam maneiras de minimizar o *trade-off* entre privacidade e utilidade na divulgação de dados quando estes passam por processo de proteção.

Os resultados deste trabalho mostraram que, o uso das operações de supressão e generalização, pode ser uma alternativa para minimizar a possível correlação com dados divulgados em outras bases de dados e manter a mesma representação semântica e veracidade dos dados.

Adicionalmente, os procedimentos de adição de ruídos e de transformada da distância associados aos métodos de agrupamento mostraram a similaridade entre os dados originais e os dados modificados, indicando uma maior utilidade e minimizando as violações de privacidade.

Portanto, as técnicas de anonimização demonstradas neste trabalho podem oferecer uma alternativa para a divulgação dos dados, minimizando possíveis ameaças à privacidade dos sujeitos envolvidos em consultas médicas e, ao mesmo tempo, oferecendo informações à sociedade e aos pesquisadores.

Artigo recebido em 03/07/2016 e aceito para publicação em 23/02/2017

ANALYSIS OF THE BALANCE BETWEEN PRIVACY AND UTILITY IN DATA ACCESS

ABSTRACT *This article analyzes measures for privacy protection by means of non-perturbative and perturbative techniques, aiming equilibrium between privacy and utility of data. Then, suppression and generalization were used for anonymization of medical consultation data and, based on Mivule (2015) the addition of noises was performed to generate private modified data. Subsequently, the utility of those data was verified by means of cluster analysis. The results showed the generalization preserves the accuracy of the data and, when combined with the addition of noises, can be a viable alternative to disclose data, minimizing risks to privacy and with an acceptable level of utility.*

Keywords: *Data protection. Anonymization. Data utility.*

REFERÊNCIAS

ADAMS, A.; SASSE, M. A. Privacy in Multimedia Communications: Protecting Users, Not Just Data. In: BLANDFORD, A. (Org). **People and computers XV – Interaction without frontiers.** [S.l.]: Springer London, p. 49-64, 2001. Disponível em: <http://www.eis.mdx.ac.uk/ridl/aadams/hci01.pdf>. Acesso em: 05 dez. 2015.

ANS**.** Padrão para Troca de Informação de Saúde Suplementar - TISS. In: **Agência Nacional de Saúde Suplementar**. Disponível em: <http://www.ans.gov.br/prestadores/tiss-troca-de-informacao-de-saude-suplementar>. Acesso em: 20 jun. 2016.

BETTINI, C.; RIBONI, D. Privacy protection in pervasive systems: State of the art and technical challenges. **Pervasive and Mobile Computing**, v. 17, p. 159-174, 2015.

BORKO, H. Information science: what is it?. **American documentation**, v. 19, n. 1, p. 3-5, 1968.

CIRIANI, V. *et al*. Microdata protection. In: YU, T. JAPODIA, S. (Org.). **Secure data management in decentralized systems**. [S.l.]: Springer US, 2007. p. 291-321.

CIRIANI, V. *et al*. Theory of privacy and anonymity. In: ATALLAH, M.; BLANTON, M. (Org). **Algorithms and theory of computation handbook.** Chapman & Hall/CRC, 2010. p. 18-18.

CORRAR, L. J.; PAULO, E.; DIAS FILHO, J. M. **Análise multivariada: para os cursos de administração, ciências contábeis e economia**. São Paulo: Atlas, p. 280-323, 2009.

DAVIES, D.; BOULDIN, D. W. A cluster separation measure. In: FORSYTH D. A. (Org.). **IEEE transactions on pattern analysis and machine intelligence**, n. 2, p. 224-227, 1979. Transactions on, n. 2, p. 224-227, 1979.

DE CAPITANI DI VIMERCATI, S. *et al*. Data privacy: Definitions and techniques**. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems**, v. 20, n. 06, p. 793-817, 2012. Disponível em: <http://spdp.di.unimi.it/papers/ijufks2012.pdf>. Acesso em: 17 dez. 2015.

DOMINGO-FERRER, J.; SEBÉ, F.; CASTELLÀ-ROCA, J. On the security of noise addition for privacy in statistical databases. In: DOMINGO-FERRER, J; TORRA, V. (Org.). [S.l.]: **International Workshop on Privacy in Statistical Databases**. Springer Berlin Heidelberg, 2004. p. 149-161.

DWORK, C. Differential privacy: A survey of results. In: **International Conference on theory and applications of models of computation**. Springer Berlin Heidelberg, 2008. p.1-19.

FUNG, B.C.M. *et al.* **Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques** (1st ed.). [S.l.]: Chapman & Hall/CRC, 2010.

KARGUPTA, H. et al. Random-data perturbation techniques and privacy-preserving data mining. **Journal Knowledge and Information Systems**, v.7, n. 4, p. 387-414, 2005.

KIM, J. J. A method for limiting disclosure in microdata based on random noise and transformation. In: **Proceedings of the section on survey research methods. American Statistical Association**, 1986. p. 303-308.

LE COADIC, Y. F. **A Ciência da Informação**. Tradução de Maria Yêda F. S. de Filgueiras Gomes. Brasília: Briquet de Lemos, 1996.

MARTINS, G. A. M. **Estatística Geral e Aplicada**. São Paulo: Atlas, 2002. 417p.

MIVULE, K. **On the Generation of Privatized Synthetic Data Using Distance Transforms**. 2015. Disponível em: <https://www.researchgate.net/profile/Kato_Mivule/publication/274030996_On_the_Generation_of_Privatized_Synthetic_Data_Using_Distance_Transforms/links/5512b8fd0cf20bfdad51d8c3.pdf>. Acesso em: 04 nov. 2015.

_____. **An investigation of data privacy and utility using machine learning as a gauge**. 2014. Tese de Doutorado. BOWIE STATE UNIVERSITY. Disponível em: <https://eric.ed.gov/?id=ED569049>. Acesso em: 28 jan. 2016.

_____. **Utilizing noise addition for data privacy, an overview**. arXiv preprint arXiv:1309.3958, 2013. Disponível em: <http://arxiv.org/abs/1309.3958>. Acesso em: 19 out. 2015.

MORETTIN, P. A.; TOLOI, C. **Análise de séries temporais**. São Paulo: Edgard Blucher, 2004. 556 p.

NERGIZ, M. E.; GÖK, M. Z. Hybrid k-Anonymity. **Computers & Security**, v. 44, p. 51-63, 2014.

RUN, C. et al. Protecting Privacy Using K-Anonymity with a Hybrid Search Scheme. **International Journal of Computer and Communication Engineering**, v. 1, n. 2, p. 155, 2012.

SAMARATI, P.; SWEENEY, L. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. **Technical report, SRI International**, 1998. Disponível em: <https://epic.org/privacy/reidentification/Samarati_Sweeney_paper.pdf>. Acesso em: 05 nov. 2015.

SÁNCHEZ, D.; BATET, M. C-sanitized: A privacy model for document redaction and sanitization. **Journal of the Association for Information Science and Technology**, 2015. Disponível em: <http://arxiv.org/ftp/arxiv/papers/1406/1406.4285.pdf> . Acesso em: 10 jan. 2016.

SKOPEK, J. M. Anonymity, the production of goods, and institutional design. **Fordham L. Rev.**, v. 82, p. 1751, 2014. Disponível em: <http://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=4960&context=flr>. Acesso em: 07 nov. 2015.

WESTIN, A. F. **Privacy and Freedom**. New York: Ig Publishing. 1967. 558 p.

WONG, R. Chi-Wing et al. (α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing. In: **Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining**. ACM, 2006. p. 754-759.