PRESERVAÇÃO DA PRIVACIDADE NO ACESSO A DADOS POR MEIO DO MODELO K-ANONIMATO

Elaine Parra Affonso
Universidade Estadual Paulista - Unesp/Marília
Doutoranda em Ciência da Informação
elaineaffonso@marilia.unesp.br

Ricardo César Gonçalves Sant'Ana
Universidade Estadual Paulista - Unesp/Marília
Doutor do Programa de Pós-Graduação em Ciência da Informação – Unesp/Marília – SP
ricardosantana@marilia.unesp.br

Resumo: O grande desafio para as organizações é garantir a preservação da privacidade ao disponibilizar dados sensíveis, pois corre-se o risco de que seja obtida correlação dos dados privados com base de dados pública, o que pode levar a quebra de confidencialidade. O objetivo deste artigo é demonstrar que existem meios de minimizar problemas relacionados à divulgação de dados sensíveis. Por meio da estrutura de dados disponibilizada no padrão TISS – Troca de Informação em Saúde Suplementar, foi simulada uma base de dados que recebeu generalização e supressão, operações do modelo K-anonimato. Posteriormente foram efetuados ataques, identificando possíveis vulnerabilidades na base de dados, com a finalidade de validar o processo de anonimização. A retirada dos identificadores não é suficiente para atingir o anonimato, pois ao combinar atributos de base de dados privada com públicas é possível à revelação de informações confidenciais, inclusive o atacante pode utilizar-se do conhecimento prévio e correlacionar com os dados disponíveis, principalmente quando a quantidade de semi-identificadores é expressiva na tabela de dados. Com o aumento na coleta e compartilhamento de dados, conjuntamente com a necessidade de acesso, torna-se relevante o estudo e a análise dos aspectos que implicam na disponibilização dos dados e na preservação da privacidade.

Palavras-chave: Anonimato. Dados sensíveis. Semi-identificadores. Identificadores pessoais.

PRIVACY PRESERVATION IN ACCESS TO MODEL THROUGH DATA K- ANONYMITY

Abstract: The major challenge for organizations is to ensure the privacy preservation by providing sensitive data, as it runs the risk that a correlation of private data based on public data, is obtained, leading to a breakdown in the confidentiality. The purpose of this article is to demonstrate there are ways to minimize problems related to the disclosure of sensitive data. Through the data structure available in standard TISS - Information Exchange Standard, a database that received generalization and suppression operations stipulated by the K-anonymity model was simulated. Subsequently attacks were performed, identifying possible vulnerability in the database, in order to validate the anonymizing process. The removal of the identifiers is not enough to achieve anonymity because when combining attributes present in a private database with public, may cause to the disclosure of confidential information, moreover, the attacker may use the prior knowledge correlate to the available data, mainly when the amount of semi-identifiers is significant in the data table. With the increase in the collection and sharing of data, together with the need to allow them the access, it brings the necessity of a study and analysis of the aspects that involve the data provision and the privacy preservation.

Keywords: Anonymity. Sensitive data. Semi-identifiers. Personal identifiers.

1 INTRODUÇÃO

Por intermédio das Tecnologias da Informação e Comunicação (TIC) torna-se inevitável a coleta e o armazenamento de dados dos mais diversos segmentos. Na maioria das vezes o usuário não tem a percepção do destino desta coleta, inclusive o tempo de armazenamento dos seus dados. De modo geral, o usuário não tem garantia e opção de negociar que seus dados privados ou combinados a assuntos de interesses pessoais ou sigilosos sempre serão utilizados de modo ético.

Para Sweeney (2002) a sociedade está experimentando um crescimento significativo na quantidade e variedade de dados contendo informações pessoais dos indivíduos, adquiridas decorrentes o uso das TIC. Os responsáveis pelos dados muitas vezes apresentam dificuldades na disponibilização de informações que não comprometa a privacidade e confidencialidade. Existem situações que a utilidade do banco de dados depende da capacidade do responsável em produzir dados anônimos, pois muitas vezes o ato de não publicar tais informações pode diminuir a importância dos dados, enquanto que por outro lado, não fornecer a proteção adequada, pode criar situações que prejudiquem o indivíduo, revelando dados de contexto pessoal.

A crescente demanda de acesso a dados e informações por meio das tecnologias implica na necessidade do aumento de recursos e melhorias em todas as fases do ciclo de vida dos dados, desde a coleta até a sua disponibilização, e cabe a Ciência da Informação o papel de desempenhar esta importante tarefa (SANT'ANA, 2013), pois esta "relaciona com o corpo de conhecimentos relativos à produção, coleta, organização, armazenamento, recuperação, interpretação, transmissão, transformação e utilização da informação" (BORKO, 1968, pg. 3, tradução nossa).

A Ciência da Informação tem um importante papel na construção de um arcabouço teórico que atenda a demanda pela busca de informações, uma vez que é papel da Ciência da Informação investigar o comportamento da informação, bem como o fluxo e o acesso a esta (BORKO, 1968).

A informação é considerada condição básica para o desenvolvimento econômico juntamente com o capital, o trabalho e a matéria-prima, mas o que torna a informação importante atualmente é a sua natureza digital (CAPURRO; HJORLAND 2003, tradução nossa).

Devido à natureza digital da informação conjuntamente com a necessidade de disseminação e acesso aos dados, torna-se imprescindível um maior comprometimento nos aspectos que envolvam a preservação da privacidade do sujeito.

Este trabalho justifica-se pela importância de prover informações para que os detentores de dados possam garantir o anonimato ao disponibilizar dados. A preservação do anonimato tornou-se uma questão importante, pois cada vez mais as organizações têm que disponibilizar dados para o público e quando estes dados podem ser combinados com outras bases de dados é necessário utilizar alguma técnica para atingir o anonimato (Run et al., 2012; Wong et al., 2006).

Neste contexto, é relevante refletir sobre os seguintes aspectos: É possível proteger as questões pessoais e ao mesmo tempo permitir o acesso aos dados? Um detentor de dados pode disponibilizar dados privados com garantia de que os sujeitos não poderão ser identificados e manter a utilidade dos dados, atendendo as necessidades do público alvo?

Zimbardo (1969 apud CHRISTOPHERSON, 2007) definiu o anonimato como a incapacidade de identificar o indivíduo de tal forma que ele não possa ser avaliado, criticado, julgado ou punido. O anonimato não significa que necessariamente uma pessoa não é visível para outros, mas é necessário que os outros sejam incapazes de identificar essa pessoa como um indivíduo. Pode-se associar o anonimato a visão de Foucault "É visto, mas não vê, objeto de uma informação, nunca sujeito numa comunicação" (FOUCAULT, 1987, p. 224).

O objetivo geral deste artigo é analisar o k-anonimato como alternativa para minimizar quebras de privacidade na disponibilização de dados, utilizando os elementos do padrão TISS – Troca de Informação em Saúde Suplementar, disponibilizados pela ANS – Agência Nacional de Saúde.

A ANS é a agência reguladora pertencente ao Ministério da Saúde que é responsável pelo setor de planos de saúde no Brasil, e para possibilitar a interoperabilidade dos dados entre as operadoras de plano de saúde e estabelecimentos, utiliza-se o padrão TISS, padrão obrigatório para as trocas eletrônicas de dados dos beneficiários de planos de saúde e entre os agentes da saúde suplementar, com o objetivo de possibilitar a interoperabilidade funcional e semântica entre os diversos sistemas independentes, para fins de avaliação da assistência à saúde (ANS, 2015).

Espera-se com este trabalho, demonstrar que ao utilizar operações de anonimização é possível disponibilizar o acesso a dados, de forma a contribuir com pesquisadores e demais interessados, sem que ocorra a quebra de privacidade.

1.2 MODELO PARA ANONIMIZAÇÃO DE DADOS (K-ANONIMATO)

O termo anonimato representa o fato do sujeito não ser unicamente caracterizado dentro de um conjunto de sujeitos. Neste caso, afirma-se que o conjunto está anonimizado. O conceito de sujeito refere-se a uma entidade ativa, como uma pessoa ou um computador. A preocupação está em impedir a identificação de uma entidade a partir de uma análise semântica de um conteúdo, o anonimato não se destina a proteger apenas a identidade de um sujeito, mas exige que outros usuários sejam incapazes de determinar a identidade de um individuo quando combinado a uma situação ou contexto (PACHECO, 2013; PFITZMANN; KÖHNTOPP, 2001).

O anonimato pode ter um conceito multifacetado, e pode ser tratado de forma diferente dependendo do contexto. Por exemplo, os ambientes altamente controlados, como o de uma rede militar, normalmente não deseja o anonimato, enquanto outros, como uma sala de bate-papo, pode pelo menos, permitir um nível aceitável de anonimato para os seus usuários (KAMBOURAKIS, 2014).

> Existe uma diferença importante entre a privacidade e o anonimato: sob a condição de privacidade, pode-se ter o conhecimento da identidade de uma pessoa, mas não de um fato pessoal associado a ela, que, nos termos de condição do anonimato, tem-se o conhecimento de um fato pessoal, mas não da identidade da pessoa associada. Neste sentido, a privacidade e o anonimato são faces opostas uma da outra. Enquanto a privacidade, muitas vezes esconde fatos sobre alguém cuja identidade é conhecida, removendo informações e outros bens associados à pessoa de circulação pública, o anonimato, muitas vezes esconde a identidade de alguém sobre quem os fatos são conhecidos, com a finalidade de colocar tais dados em disponibilização (SKOPEK; 2014, p. 1755, tradução nossa).

Para Burkell (2006) o anonimato é definido considerando três aspectos distintos:

a) proteção da identidade: Retenção na fonte do nome ou outros identificadores únicos, o sujeito deve permanecer não identificado e está relacionado a uma entidade do mundo real;

b) anonimato de ação: É definido quando as ações do sujeito não podem ser vistas e não estão disponíveis para os outros;

c) anonimato visual: É alcançado quando o rosto do sujeito passa despercebido ou invisível para os outros.

A proteção da identidade de um indivíduo vem do aumento da coleta de um tipo específico de informações: as denominadas “informações pessoais identificadas”, conhecida pela sigla PII (*Personally Identifiable Information*). Cada entidade é descrita por atributos

ligados a ele (nome, biologia, características sociais, competências, localização, personalidade) que são aquelas referentes à vida das pessoas, incluindo desde suas características físicas até seus hábitos pessoais, sendo possível criar um perfil completo com a combinação destes dados (JENNINGS, 2000; KAMBOURAKIS, 2014).

As PII podem ser definidas como qualquer dado ou informações que são disponibilizadas pela internet e que possa ser combinada, de alguma forma, a uma pessoa física, a alguém que tem um nome, um endereço e uma vida (JENNINGS, 2000).

Para Nergiz e Gok (2014) estes dados tem um valor inestimável não só em relação à pesquisa, mas também para as perspectivas do negócio, pois pode ser uma fonte valiosa de informações para investigação médica, análise de tendências. Entretanto, se o indivíduo pode ser unicamente identificável, então os seus dados privados acabam sendo divulgados (MACHANAVAJJHALA; GEHRKE; KIEFER, 2007).

A importância destas PII pode ser representada no caso da Anthem (segunda maior companhia de seguros de saúde), onde cerca de 80 milhões de clientes tiveram suas informações de conta roubada, os hackers conseguiram acesso ao sistema da Anthem e aos dados pessoais como nomes, aniversários, identificação de médicos, número de segurança social, endereços de e-mail e informações de emprego. Os atacantes não estavam interessados em informações médicas sobre os clientes, a informação de identificação pessoal tem sido mais valiosa do que as ocorrências médicas (WEISE, 2015).

Segundo Run et al. (2012), em alguns casos, as organizações têm que disponibilizar os dados para o público, para usos especiais, tais como análise estatística ou estado de saúde do paciente para fins de pesquisa. Com o objetivo de proteger o indivíduo e garantir privacidade, os identificadores únicos, tais como nome e número de RG devem ser removidos, pois existe a possibilidade de combinar os dados com outros atributos externos para identificar os indivíduos.

Sweeney (2002) também corrobora que é uma prática comum nas organizações liberar e receber dados com todas as PII removidas como nome, endereço e número do telefone, presumindo–se que o anonimato é mantido, pois os dados resultantes desta supressão gera um olhar anônimo. No entanto, na maior parte destes casos, os dados restantes podem ser utilizados para identificar indivíduos combinando com dados de outras bases.

É possível um atacante combinar dados públicos com outros dados publicamente disponíveis de uma base privada. Esta situação tem gerado preocupações em diversos segmentos, pois os dados estão sendo cada vez mais disponibilizados para acesso, como o uso

em pesquisas, e pode estar sujeitos ao comprometimento da privacidade. Por exemplo, é desejável que se consiga revelar informações médicas de forma que as identidades dos indivíduos não possam ser reveladas (SAMARATI, 2001).

Os dados são normalmente armazenados em uma única relação R, definida por um esquema relacional R (a_1, a_2, a_3,...,a_n), onde a_1 é um atributo no domínio D_i, com i = 1,..n. Na perspectiva da divulgação de dados dos indivíduos, os atributos em R podem ser classificados da seguinte forma (CAMENISCH; FISCHER-HÜBNER; RANNENBERG,2011; SAMARATI, 2001):

a) **Identificadores (I):** atributos que identificam unicamente os indivíduos (ex.: CPF, Nome, Número da Identidade);

b) **Semi-identificadores (SI):** atributos que podem ser combinados com dados externos e expor o indivíduo, ou ainda reduzir a incerteza sobre suas identidades (ex.: data do nascimento, CEP, cargo, função, tipo sanguíneo);

Fung et al. (2010) e Run et al. (2012) corrobora com a ideia de que os semi-identificadores é um conjunto de atributos que quando combinados podem identificar o registro.

Para exemplificar os semi-identificadores Sweeney (2002) relata a seguinte situação: Dado um conjunto de elementos em uma entidade U, na entidade-especificada T (A_1,...,A_n), f_c:U→T e f_g: T→U', onde U⊆U'. O semi-identificador de T, escrito Q_T, é o conjunto de atributos $\{A_1,...,A_j\} \subseteq \{A_1,...,A_n\}$ onde $\exists pi \in U$ ambos que $f_g(f_c(p_i)[Q_T])=p_i$.

c) **Atributos sensíveis (AS)**: ou também chamados de confidenciais, representam os atributos que contêm informações sensíveis sobre os indivíduos (ex.: doenças, salário, exames médicos, lançamentos do cartão de crédito) (VIMERCATI et al., 2012).

Na literatura de Fung et al (2010) e Vimercati et al. (2012), os autores citam também os atributos não sensíveis, caracterizados como aqueles que contém todos os atributos que não se enquadram nas três categorias anteriores e a sua divulgação não causa nenhum tipo de problema.

Desta forma, a preservação da privacidade tornou-se uma questão importante na divulgação dos dados, quando um conjunto de dados é disponibilizado, alguma técnica de preservação da privacidade faz-se necessária para minimizar a possibilidade de identificar informações sensíveis sobre os indivíduos (WONG et al., 2006).

> O acesso à base de dados publicada não deveria disponibilizar ao adversário o conhecimento da informação extra sobre qualquer assunto da vitima e nenhum acesso à base de dados, mesmo que o adversário já possua informações do individuo obtidas anteriormente de outras fontes. (DALENIUS, 1977 apud FUNG et al., 2010, p. 13)

A consequência da disseminação da informação pode permitir que esta seja compreendida por um atacante, que se refere a qualquer pessoa com legítimo acesso aos dados.

> Assim, este compartilhamento de dados levanta preocupações de privacidade e devem ser anonimizados antes de serem publicados, que deve não apenas satisfazer as exigências de privacidade, mas também a utilidade dos dados (NERGIZ; GOK, 2014, p 1. tradução nossa).

As principais operações para anonimizar dados e minimizar perdas de privacidade incluem: generalização, supressão, permutação e perturbação de dados (Fung et al., 2010; VIMERCATI et al., 2012). "Uma característica comum destes métodos é que eles manipulam os dados utilizando a generalização" (NERGIZ; GOK, 2014, tradução nossa).

A generalização é usada para enfatizar as semelhanças entre tipos de entidades de nível superior e ocultar suas diferenças (KORTH; SILBERSCHATZ, 1993). Para Nergiz e GoK (2014) a generalização consiste em substituir valores de dados com valores gerais (valores que incluem o significado original e que podem também implicar outros valores, por exemplo, "Itália" é alterado para "Europa"), de modo que mais registros possam expressar significados semelhantes.

Wong et al. (2006) também corrobora que cada atributo tem um correspondente de hierarquia conceitual, onde um domínio de nível mais baixo na hierarquia fornece mais detalhes do que um domínio de nível superior. Por exemplo, data de nascimento dia/mês/ano (15/03/2015) é um domínio de nível inferior e data de nascimento Y (por exemplo, 2015) é um domínio de nível superior.

Para CIRIANI et al. (2007) a generalização pode ser aplicada em relação a:

a) Atributo: a generalização ocorre nos valores da coluna;
b) Células: a generalização é realizada em células individuais, como resultado, a tabela pode conter em uma coluna específica, valores distintos nos atributos.

Esta técnica substitui os valores de atributos semi-identificadores por valores menos específicos, conseguindo assim tornar o dado anônimo, mas semanticamente consistentes.

A vantagem da generalização é que ao contrário das outras técnicas para garantir o anonimato, ela preserva a veracidade das informações e diminui o número de registros diferentes em uma tabela de dados (SAMARATI, 2001).

Outra abordagem para conseguir realizar a preservação da privacidade é a supressão, que consiste em suprimir os dados de uma tabela de maneira que eles não sejam disponibilizados (SAMARATI, 2001).

Para Ciriani et al. (2007), a supressão pode ocorrer de três maneiras:

a) no registro: onde a supressão é efetuada na linha, a supressão remove todo o registro;

b) na coluna: a supressão elimina todos os valores de uma coluna;

c) nas células individuais: podem ser supridas apenas algumas células (atributos) de um determinado registro.

O objetivo dos métodos para preservar a privacidade e atingir o anonimato é essencialmente evitar que a divulgação dos dados de contexto possam ser combinados por um atacante para re-identificar os sujeitos na base de dados.

Este trabalho demonstra os aspectos envolvidos na preservação da privacidade por meio do modelo k-anonimato.

> O k-anonimato institui uma fundamentação teórica formal para o problema da quebra de privacidade na questão de divulgação de microdados, anonimizados com supressão ou generalização de dados, estabelecendo como o anonimato pode ser garantido no nível desejado aplicando a generalização ou supressão apropriadas (PACHECO, 2013, p.66).

O modelo K-anonimato é uma das maneiras mais populares para resolver o problema da proteção da privacidade (RUN et al., 2012). Está técnica de anonimização de dados generaliza valores e/ou suprime registros, com o objetivo de garantir que cada registro procurado possa ser associado com pelo menos k possíveis correspondentes, o valor de K é representado por um valor inteiro positivo, definido pelo proprietário dos dados. Quanto maior for o valor numérico de K, maior será anonimização e consequentemente menor o risco de identificação do indivíduo, pois a probabilidade de re-identificar um registro é de 1/k, embora isto não proteja a base de dados contra divulgação de atributos sensíveis, pois mesmo que o atacante não tenha capacidade de re-identificar o registro, ele pode descobrir atributos sensíveis na base de dados anonimizada (BETTINI; RIBONI, 2014; RUN et al, 2012; SAMARATI; SWEENEY, 1998).

A teoria do K-anonimato obteve muito reconhecimento ao demonstrar matematicamente que ao estratificar “dados gerais” é possível dificultar a rastreabilidade e a combinação de dados. Os atributos estratificados foram chamados de variável K e ao ter alguns atributos “k”, chama-se de “k-anonimato”, quanto maior for o valor numérico de “k”, maior será a anonimização (GÖRLICH, 2015).

O k-anonimato por meio das operações de generalização e supressão propõe uma abordagem para proteger a identidade do indivíduo enquanto mantém os valores dos atributos

verdadeiros. A questão é saber se é melhor generalizar, perdendo a precisão de dados, ou suprir e perder a completude (CIRIANI et al., 2007).

Embora o método do K-anonimato tem o objetivo de proteger os dados, acaba sendo vulnerável a alguns ataques que podem levar a violação de privacidade, por exemplo, a técnica não protege os atributos sensíveis de serem descobertos quando um grupo de semi-identificadores não possui valores diferentes nos atributos sensíveis. Sob esta premissa, acontece o ataque de homogeneidade, quando um atacante sabe o valor do semi-identificador de um sujeito e pode inferir a um valor sensível associado a uma pessoa (CIRIANI et al., 2007; FRIEDMAN; WOLFF; SCHUSTER, 2008).

O modelo k-Anonimato pode criar grupos que disponibilizam informações devido à falta de diversidade no atributo sensível (MACHANAVAJJHALA; GEHRKE; KIEFER, 2007).

Fung et al (2010) classifica dois tipos de ataques possíveis em base de dados: a divulgação da identidade, que ocorre quando um atacante pode identificar um sujeito a partir dos dados publicados, associando-o em um registro da tabela ou a um atributo sensível da tabela ou a tabela inteira. O segundo tipo de ataque ocorre quando o atacante tem conhecimento antes e depois de ter acesso à base de dados anonimizada, sem necessariamente associar a um registro específico.

Para Friedman, Wolff e Schuster (2008), uma limitação em relação ao modelo k-anonimato é a dificuldade para o proprietário de uma base de dados determinar quais atributos que estão disponíveis em uma tabela externa. Outra deficiência do modelo k-anonimato é quando os atacantes possuem algum conhecimento mais detalhado sobre o sujeito (MACHANAVAJJHALA; GEHRKE; KIEFER, 2007).

Para solucionar estes problemas, emerge outros modelos com o objetivo de possibilitar a anonimização de dados: como l-diversity (MACHANAVAJJHALA; GEHRKE; KIEFER, 2007), LKC-Privacy (MOHAMMED; FUNG; DEBBABI, 2010); t-closeness (LI; LI; VENKATASUBRAMANIAN, 2007); b-likeness (CAO; KARRAS, 2012), entre outros, que não foram objeto de estudo deste trabalho.

Apesar destas limitações, o k-anonimato é um dos modelos mais aceitos para preservação da privacidade, por várias razões: o modelo k-anonimato define a privacidade da saída de um processo e não do próprio processo; é um método simples, intuitivo e permite que os responsáveis pelos dados tenham a certeza que estão utilizando o modelo corretamente (FRIEDMAN; WOLFF; SCHUSTER, 2008).

2 METODOLOGIA

Foi utilizado o modelo K-anonimato (SWEENEY, 2002), considerando a aplicação da generalização de atributos e supressão de registros.

O modelo foi demonstrado em uma tabela com dados simulados de sujeitos envolvidos em consultas médicas. A tabela foi estruturada a partir dos elementos da guia de consulta definidas no padrão TISS, disponibilizado pela ANS[1].

Este trabalhado teve como escopo a fase de recuperação do ciclo de vida dos dados (SANT'ANA, 2013) dos prestadores de serviços de saúde, considerando a estrutura dos dados utilizados na guia de consulta.

A primeira etapa consistiu em classificar os atributos da guia de consulta como identificadores, semi-identificadores e sensíveis; seguido da supressão dos atributos que são considerados únicos na tabela (identificadores).

A segunda etapa simulou especificamente os ataques por atributo, por registro e por homogeneidade de dados na tabela.

Posteriormente utilizou as operações definidas pelo modelo k-anonimato (generalização e supressão), verificando o grau do anonimato na tabela proposta. Novos ataques foram efetuados, buscando identificar eventuais fragilidades e como forma de validar o processo de anonimização.

3 RESULTADOS E DISCUSSÕES

O objetivo de realizar a anonimização dos dados é minimizar problemas decorrentes da divulgação de dados do contexto do sujeito quando estes permitem ser combinados com outras bases de dados, tornando-se possível a disponibilização dos dados que possam ser úteis para a sociedade sem inferir na quebra de privacidade.

A descrição dos atributos e condição de preenchimento disponibilizadas no Quadro 1 foi base para a demonstração das operações do k-anonimato (generalização e supressão) simuladas neste trabalho.

[1]http://www.ans.gov.br/prestadores/tiss-troca-de-informacao-de-saude-suplementar/padrao-tiss-versao-3-02-01

Quadro 1: Estrutura de atributos da guia consulta

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
NumCart	Nº da carteira do beneficiário na operadora.	Obrigatório
RN	Indica se o paciente é um recém-nato que está sendo atendido no contrato 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 a 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 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
Conselho	Número 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 (2015)

Estruturou-se uma base de dados com valores simulados (Quadro 2) representado os dados coletados na guia de consulta, na qual foi aplicada as operações de anonimização que constituem o modelo K-anonimato.

Quadro 2: Representação dos dados simulados das guias de Consulta

RegANS	NGuiaPrest	NGuiaOp	NumCart	ValCart	RN	Benef	CNS	CodOp	NomeContr	CNES	NomeProf	CProf	Conselho	UF	CBO	IndAc	DtAtend	Con	Tab	CodProc	VlProc	OBS
426736	1234567	987654	44261009	12/02/2016	não	Clara Shutz	12345	100000	Hosp. São Mateus	3909878	Luiza Ventura	6	578934	SP	225250	1	12/02/2015	3	22	10101012	0	-
426736	1234568	987653	44261008	11/05/2015	não	Cássia Cabral	12346	200000	Hosp. Sta Luzia	2080515	Ademar Caetano	6	388081	SP	225250	9	13/02/2015	4	22	10106049	0	-
426736	1234570	987652	44261007	30/07/2017	não	Joana Silva	12347	300000	Hosp. Regional	3909866	Luiza Ventura	6	578934	SP	225250	9	20/02/2015	1	22	10101012	0	-
326310	1234571	987734	44261006	24/08/2015	não	Miguel Luiz	12348	200000	Hosp. Sta Luzia	2080515	Roberto Souza	6	456789	SP	225121	0	14/03/2015	1	22	10101012	0	-
426736	1234572	987651	44261005	10/02/2016	não	Pedro Santos	12351	200000	Hosp. Sta Luzia	2080515	Vitor Dias	6	345678	SP	225121	9	29/03/2015	1	22	10101012	0	-
426736	1234573	987401	44261001	12/09/2015	não	Luiz Costa	12352	200000	Hosp. Sta Luzia	2080515	José Mauá	6	123456	SP	225175	9	14/02/2015	1	22	10101012	0	-
426736	1234574	987402	44261004	30/07/2017	não	Laís Val	12353	200000	Hosp. Sta Luzia	2080515	Luiza Ventura	6	578934	SP	225250	9	20/02/2015	1	22	10101012	0	-
326310	1234575	987733	44261003	03/04/2016	não	Mario Soares	12354	200000	Hosp. Sta Luzia	2080515	Roberto Souza	6	456789	SP	225121	9	19/03/2015	1	22	10101012	0	-

Fonte: Elaborado pelos autores

Ao disponibilizar esses dados, é possível colaborar, por exemplo, com pesquisadores, ao prover o tipo de consulta, ocorrência de acidentes e procedimentos. Todavia, tornar útil esta disponibilização e acesso pode comprometer a privacidade do indivíduo, mesmo que os dados pessoais sejam omitidos. A proposta do k-anonimato é mitigar ataques à privacidade do sujeito, considerando a possibilidade do atacante combinar dados privados com outras tabelas públicas ou pelo conhecimento adquirido anteriormente.

3.1 PROCESSO DE ANONIMIZAÇÃO: CLASSIFICAÇÃO DOS ATRIBUTOS

O primeiro passo no modelo K-anonimato proposto por Sweeney (2002) é classificar os atributos identificadores, semi-identificadores e sensíveis (Quadro 3).

Quadro 3: Classificação dos atributos da guia de consulta

Tipo do Atributo	Atributo	Atributos da Guia Consulta
Identificador	NGuiaPrest	Nº que identifica a guia no prestador de serviços
	NGuiaOp	Número da guia atribuído pela operadora
	NumCart	Número da carteira
	Benef	Nome do beneficiário
	CNS	Número do CNS
Semi-Identificadores	RegANS	Registro da operadora de plano privado na ANS
	ValCart	Validade da carteira
	CodOp	Código identificador do prestador executante junto à operadora,
	NomeContr	Nome do contratado
	CNES	Código do CNES
	NomeProf	Nome do profissional executante
	Cons	Número do conselho
	CBO	Código CBO
Atributos Sensíveis	IndAc	Tipo identificador de acidente
	Con	Tipo de consulta
	Tab	Tabela
	CodProc	Código do procedimento

Fonte: Elaborado pelos autores

No quadro 3 foi identificado os atributos que identificam unicamente o indivíduo (identificador), os atributos que podem aparecer tanto em uma base de dados privada quanto em uma base de dados pública (semi-identificadores) e os atributos que contêm informações pessoais da vida do indivíduo (atributos sensíveis).

Os atributos atendimento ao RN, conselho profissional, UF, valor do procedimento e observação, caracterizam-se como atributos não sensíveis, pois não se enquadram nas três categorias anteriores. Embora o atributo observação possa ter qualquer valor, este não é estruturado, portanto, dificulta os ataques e caracteriza-se como uma vantagem em termos de anonimato.

Após definir os atributos identificadores, estes foram removidos da tabela (Quadro 4), este é o primeiro passo para a anonimização, a supressão dos identificadores únicos. Foi suprido da tabela o conjunto de atributos: número de identificação da guia no prestador de serviço, número da guia atribuído pela operadora; número da carteira; nome do beneficiário e número do CNS.

Quadro 4: Resultado após a supressão dos atributos identificadores

RegANS	ValCart	RN	CodOp	NomeContr	CNES	NomeProf	CProf	Conselho	UF	CBO	IndAc	DtAtend	Con	Tab	CodProc	VlProc	OBS
426736	12/02/2016	não	100000	Hosp. São Mateus	3909878	Luiza Ventura	6	578934	SP	225250	1	12/02/2015	3	22	10101012	0	-
426736	11/05/2015	não	200000	Hosp. Sta Luzia	2080515	Ademar Caetano	6	388081	SP	225250	9	13/02/2015	4	22	10106049	0	-
426736	30/07/2017	não	300000	Hosp. Regional	3909866	Luiza Ventura	6	578934	SP	225250	9	20/02/2015	1	22	10101012	0	-
326310	24/08/2015	não	200000	Hosp. Sta Luzia	2080515	Roberto Souza	6	456789	SP	225121	0	14/03/2015	1	22	10101012	0	-
426736	10/02/2016	não	200000	Hosp. Sta Luzia	2080515	Vitor Dias	6	345678	SP	225121	9	29/03/2015	1	22	10101012	0	-
426736	12/09/2015	não	200000	Hosp. Sta Luzia	2080515	José Mauá	6	123456	SP	225175	9	14/02/2015	1	22	10101012	0	-
426736	30/07/2017	não	200000	Hosp. Sta Luzia	2080515	Luiza Ventura	6	578934	SP	225250	9	20/02/2015	1	22	10101012	0	-
326310	03/04/2016	não	200000	Hosp. Sta Luzia	2080515	Roberto Souza	6	456789	SP	225121	9	19/03/2015	1	22	10101012	0	-

Fonte: Elaborado pelos autores

Caso alguns dos atributos identificadores fossem disponibilizados, como por exemplo, o atributo CNS, e o atacante tivesse conhecimento de alguns dados pessoais do sujeito, é possível por meio do acesso ao site "consulta cartão SUS"[2] que solicita o CPF, o município de nascimento e o mês de nascimento, descobrir o CNS do sujeito e correlacionar com os dados disponibilizados (Quadro 2), obtendo conhecimento das ações deste sujeito.

Os atributos nome do beneficiário, número da carteira, número da guia atribuído pela operadora e número da guia no prestador foram supridos porque são valores únicos para cada guia, permitindo uma identificação do paciente com o contexto da guia.

[2] Disponível em http://cartaosus.com.br/consulta-cartao-sus/

Para demonstrar a fragilidade de suprir apenas os identificadores, simulou-se um ataque ao atributo (Figura 1). Caso o atacante saiba que a beneficiária Clara Schutz foi à médica Luiza Ventura no dia 12/02/2015 e, neste dia a médica fez apenas um procedimento, o atacante pode ter certeza da identificação da paciente e apropriar-se das informações relativas à consulta desta, como por exemplo, que ela sofreu um acidente, pois no atributo "indicação de acidente" está com o valor 1, que de acordo com a tabela de domínio da ANS, representa acidente de trânsito, inclusive que a Dra Luiza Ventura é uma ginecologista/obstetra (CBO), e descobre o contexto da consulta da paciente Clara Schutz.

Figura 1: Ataque de ligação ao atributo

NomeContr	CNES	NomeProf	CProf	Conselho	UF	CBO	IndAc	DtAtend	Con	Tab	CodProc
Hosp. São Mateus	3909878	Luiza Ventura	6	578934	SP	225250	1	12/02/2015	3	22	10101012

Data Atendimento	Medico
12/02/2015	Luiza Ventura

Fonte: Elaborado pelos autores

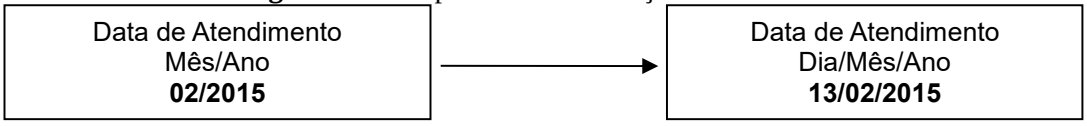
Como citado por Burkell (2006) e Nergiz e Gok (2014) a retenção dos identificadores não garante o aspecto da proteção da identidade, pois ainda permite a identificação do sujeito. Em relação ao anonimato de ação estipulado por Burkell (2006), não é atingido nesta situação (Quadro 4), pois o conteúdo e ações dos sujeitos ainda mantêm-se disponíveis.

Desta forma, observou-se a necessidade da aplicação de outras operações para garantir o anonimato dos dados, realizando assim, a generalização dos dados.

3.2 APLICAÇÃO DA GENERALIZAÇÃO

Foi utilizada a proposta idealizada por Samarati e Sweeney (1998) que é denominada de transformação dos dados por operação de generalização, que substitui os valores dos atributos semi-identificadores por valores menos específicos (Figura 2), mas permitindo a representação semântica destes dados. Para Nergiz e GoK (2014) um recurso interessante da generalização é a preservação da veracidade dos dados.

Figura 2: Exemplo de Generalização de dados



Fonte: Elaborado pelos autores

Com o objetivo de evitar a identificação do sujeito quando realizado um ataque de ligação por atributo, generalizou a coluna dos atributos que possuem valor em formato de data (Quadro 5), validade da carteira e a data de atendimento. Portanto, é observado que acontece uma recodificação global, como citado por Wong et al. (2006), pois todos os valores de um atributo provêm do mesmo nível de domínio na hierarquia, por exemplo, todos os valores na data de atendimento e validade da carteira estão no formato mês e ano.

Quadro 5: Atributos data de atendimento e validade da carteira generalizados

ValCart	RN	CodOp	NomeContr	CNES	NomeProf	CProf	Conselho	UF	CBO	IndAc	DtAtend
02/2016	não	100000	Hosp. São Mateus	3909878	Luiza Ventura	6	578934	SP	225250	1	02/2015
05/2017	não	200000	Hosp. Sta Luzia	2080515	Ademar Caetano	6	388081	SP	225250	9	02/2015
07/2017	não	300000	Hosp. Regional	3909866	Luiza Ventura	6	578934	SP	225250	9	02/2015
08/2015	não	200000	Hosp. Sta Luzia	2080515	Roberto Souza	6	456789	SP	225121	0	03/2015
02/2016	não	200000	Hosp. Sta Luzia	2080515	Vitor Dias	6	345678	SP	225121	9	03/2015

Fonte: Elaborado pelos autores

Conforme citou Nergiz e Gok (2014), "com o uso da generalização é observado o aumento dos registros que expressam significados semelhantes", ao observar o Quadro 5, é maior o número de ocorrências do mês 02 na data de atendimento.

3.2.1 Ataque no atributo

Após aplicar a generalização, e considerando o ataque por atributo simulado na Figura 1: "o atacante tem conhecimento que sujeito foi ao médico no dia 12/02/2015 e consultou-se com a médica Luiza Ventura". Considerou-se o conjunto de semi-identificadores SI={02/2015, Luiza Ventura), portanto, é possível distinguir no Quadro 6 três registros contendo os semi-identificadores e, é possível deduzir que 1/3 do conjunto de dados demonstra que a paciente está grávida, pois a consulta foi do tipo pré-natal, no atributo tipo de consulta (Con) é apresentado o valor 3, que corresponde na tabela de domínio 52 da ANS ao valor "pré-natal". Caso o atacante soubesse que a paciente se consultou no hospital São Mateus, novamente é possível concluir que o registro é da paciente Clara Shutz e descobre que ela está grávida.

Quadro 6: Registros quando realizado ataque por atributo

RN	CodOp	NomeContr	CNES	NomeProf	CProf	Conselho	UF	CBO	IndAc	DtAtend	Con	Tab	CodProc
não	100000	Hosp. São Mateus	3909878	Luiza Ventura	6	578934	SP	225250	1	02/2015	3	22	10101012
não	300000	Hosp. Regional	3909866	Luiza Ventura	6	578934	SP	225250	9	02/2015	1	22	10101012
não	200000	Hosp. Sta Luzia	2080515	Luiza Ventura	6	578934	SP	225250	9	02/2015	1	22	10101012

Fonte: Elaborado pelos autores

Observa-se neste ataque que quanto maior é o nível de informação do atacante sobre a vítima, maior será a chance de identificar unicamente o registro na tabela de dados, mesmo utilizando a operação de generalização, pois a quantidade de semi-identificadores é expressiva, aumentando a chance de identificação.

3.2.2 Ataque no registro

Para demonstrar o ataque no registro, considerou-se o conjunto de semi-identificadores SI={data de atendimento, profissional}, representado pelos valores SI= {02/2015, Ademar Caetano}, e os dados disponibilizados em uma tabela pública "Agenda Médicos: Convênio - X" (Figura 3).

Figura 3: Ataque no registro

CodOp	NomeContr	CNES	NomeProf	CProf	Conselho	UF	CBO	IndAc	DtAtend	Con
100000	Hosp. São Mateus	3909878	Luiza Ventura	6	578934	SP	225250	1	02/2015	3
200000	Hosp. Sta Luzia	2080515	Ademar Caetano	6	388081	SP	225250	9	02/2015	4
300000	Hosp. Regional	3909866	Luiza Ventura	6	578934	SP	225250	9	02/2015	1

Agenda Médicos: Convênio - X				
Médico	Paciente	Data	Horário	Hospital
Carlos Munhoz	Luiz Silva	12/02/2015	14:00	N. Senhora das Graças
Ademar Caetano	Cássia Cabral	13/02/2015	14:00	Santa Luzia
Vitoria Marlez	João Marins	14/02/2015	14:00	N. Senhora das Graças
Ademar Caetano	Flávia Loz	14/02/2015	15:00	N. Senhora das Graças
Ademar Caetano	Caique Souza	14/02/2105	16:00	Sta Casa

Fonte: Elaborado pelos autores

Neste caso, ao correlacionar uma tabela privada com dados públicos (Figura 3) acontece um ataque de ligação ao registro, pois é identificado um único registro que pertence as duas tabelas.

3.2.3 Ataque por homogeneidade ou na tabela

Caso o atacante saiba que o sujeito foi ao médico no mês de março de 2015 no hospital Santa Luzia, sendo representado pelo conjunto de semi-identificadores SI={data de atendimento, nome do contratante), com os valores SI={03/2015, Santa Luzia}, é possível confirmar a presença do registro do sujeito na tabela e descobrir suas informações (Quadro 7).

Quadro 7: Ataque por homogeneidade ou na tabela

CodOp	NomeContr	CNES	NomeProf	CProf	Conselho	UF	CBO	IndAc	DtAtend	Con	Tab	CodProc	VlProc	OBS
200000	Hosp. Sta Luzia	2080515	Roberto Souza	6	456789	SP	225121	0	03/2015	1	22	10101012	0	-
200000	Hosp. Sta Luzia	2080515	Vitor Dias	6	345678	SP	225121	9	03/2015	1	22	10101012	0	-
200000	Hosp. Sta Luzia	2080515	Roberto Souza	6	456789	SP	225121	9	03/2015	1	22	10101012	0	-

Fonte: Elaborado pelos autores

O atacante pode descobrir a especialidade do médico por meio da consulta do CBO do profissional na tabela de domínio utilizada pela ANS. Ao buscar o valor 225121, é recuperado à ocupação de oncologista, desta forma, pode-se afirmar que o sujeito se consultou com o médico Vitor Dias ou Roberto Souza e mesmo que o atacante não infira no registro do sujeito, com a repetição dos registros que possuem o valor 225121, ele pode concluir que a vítima está com suspeitas de câncer. Isto acontece devido à falta de diversidade no atributo sensível. MACHANAVAJJHALA, GEHRKE e KIEFER (2007) denomina esta situação de ataque da homogeneidade.

Como observado nos exemplos, o conhecimento que o atacante tem sobre o sujeito acaba implicando em problemas no uso do K-anonimato, mesmo realizando operações de generalização.

Nos ataques os valores relacionados à data de atendimento e ao estabelecimento de saúde, implicaram na descoberta do sujeito. Desta forma, foi aplicado novamente à generalização nos atributos do tipo data e supridos os atributos Nome do contratado, CNES e Código do prestador na operadora (Quadro 8).

Quadro 8: Generalização e Supressão de atributos

RegANS	ValCart	RN	NomeProf	CProf	Conselho	UF	CBO	IndAc	DtAtend	Con	Tab	CodProc	VlProc	OBS
426736	2016	não	Luiza Ventura	6	578934	SP	225250	1	2015	3	22	10101012	0	-
426736	2017	não	Ademar Caetano	6	388081	SP	225250	9	2015	4	22	10106049	0	-
426736	2017	não	Luiza Ventura	6	578934	SP	225250	9	2015	1	22	10101012	0	-
326310	2015	não	Roberto Souza	6	456789	SP	225121	0	2015	1	22	10101012	0	-
426736	2016	não	Vitor Dias	6	345678	SP	225121	9	2015	1	22	10101012	0	-
426736	2015	não	José Mauá	6	123456	SP	225175	9	2015	1	22	10101012	0	-
426736	2017	não	Luiza Ventura	6	578934	SP	225250	9	2015	1	22	10101012	0	-
326310	2016	não	Roberto Souza	6	456789	SP	225121	9	2015	1	22	10101012	0	-

Fonte: Elaborado pelos autores

Ao generalizar e suprir os atributos é impedido os ataques vinculados com os atributos de data e com o estabelecimento de saúde.

Para verificar se tabela garante o k-anonimato, foi analisado o grau do anonimato para a tabela guia de consulta.

3.3 DEFINIÇÃO DO GRAU DO ANONIMATO

Por conseguinte, foi verificado qual o valor de K de acordo com a fundamentação do K-anonimato, que estipula que quanto maior for o valor numérico de K, maior será anonimização e consequentemente menor o risco de divulgação.

Considerando o valor de K=2 para os semi-identificador SI {nome do profissional, data de atendimento), os conjuntos SI1={Ademar Caetano, 02/2015}, SI2={Vitor Dias, 03/2015} e SI3={José Mauá, 02/2015} serão excluídos, pois a quantidade de registros na tabela é menor do k (Quadro 9).

Quadro 9: Uso do k-anonimato (k=2)

RegANS	ValCart	RN	NomeProf	CProf	Conselho	UF	CBO	IndAc	DtAtend	Con	Tab	CodProc	VlProc	OBS
426736	2016	não	Luiza Ventura	6	578934	SP	225250	1	2015	3	22	10101012	0	-
426736	2017	não	Ademar Caetano	6	388081	SP	225250	9	2015	4	22	10106049	0	-
426736	2017	não	Luiza Ventura	6	578934	SP	225250	9	2015	1	22	10101012	0	-
326310	2015	não	Roberto Souza	6	456789	SP	225121	0	2015	1	22	10101012	0	-
426736	2016	não	Vitor Dias	6	345678	SP	225121	9	2015	1	22	10101012	0	-
426736	2015	não	José Mauá	6	123456	SP	225175	9	2015	1	22	10101012	0	-
426736	2017	não	Luiza Ventura	6	578934	SP	225250	9	2015	1	22	10101012	0	-
326310	2016	não	Roberto Souza	6	456789	SP	225121	9	2015	1	22	10101012	0	-

Fonte: Elaborado pelos autores

A técnica de supressão foi aplicada como um complemento para a generalização. A supressão pode anular alguns valores de atributos semi-identificadores ou excluir registros da tabela, como o resultado demonstrado no Quadro 10.

Foi constatado com a simulação destes ataques que os semi-identificadores revelam muitas informações a respeito dos sujeitos se forem revelados publicamente, principalmente quando acontece ocorrência única do registro na base de dados, revelando os seus dados sensíveis e a vinculação com o sujeito.

Na base de dados demonstrada no Quadro 10, ao considerar o valor de K=2 para os semi-identificadores SI {nome do profissional, data de atendimento) o semi-identificadores aparecem pelo menos duas vezes na tabela de dados.

Muitas vezes apenas o uso de generalização não é o suficiente (Quadro 5) e a supressão sozinha também não faz atingir resultados eficientes (Quadro 4), uma vez que exigiria a supressão de vários registros da base de dados, perdendo a completude da base de dados. A combinação das duas operações permite a divulgação do Quadro 10.

Quadro 10: Estrutura de dados anonimizados

RegANS	ValCart	RN	NomeProf	CProf	Conselho	UF	CBO	IndAc	DtAtend	Con	Tab	CodProc	VlProc	OBS
426736	2016	não	Luiza Ventura	6	578934	SP	225250	1	2015	3	22	10101012	0	-
426736	2017	não	Luiza Ventura	6	578934	SP	225250	9	2015	1	22	10101012	0	-
326310	2015	não	Roberto Souza	6	456789	SP	225121	0	2015	1	22	10101012	0	-
426736	2017	não	Luiza Ventura	6	578934	SP	225250	9	2015	1	22	10101012	0	-
326310	2016	não	Roberto Souza	6	456789	SP	225121	9	2015	1	22	10101012	0	-

Fonte: Elaborado pelos autores

Caso sejam realizados os mesmos ataques efetuados anteriormente é observado que não é possível identificar o sujeito na base de dados, pois os semi-identificadores que permitiam a identificação dos registros foram generalizados ou supridos da tabela, permitindo desta forma, a divulgação e acesso aos dados sem comprometer as questões de privacidade.

A vantagem de aplicar a supressão nos registros é que ele elimina apenas os registros individuais. Para Samarati (2001) é melhor suprimir mais registros do que impor mais generalizações, pois a supressão afeta os registros individuais, enquanto a generalização modifica todos os valores associados com um atributo, implicando em todos os valores associados ao registro.

As técnicas de generalização e de supressão buscam manter a veracidade e a preservação das informações, pois os dados divulgados estão sempre corretos, embora eles possam ser menos precisos na generalização ou ocultando alguns registros no caso da supressão.

4 CONSIDERAÇÕES FINAIS

Diante da necessidade de disponibilizar dados úteis para que a sociedade possa se apropriar da informação, se faz necessário que os dados sejam anônimos. O sujeito não pode ser unicamente identificado dentro de um conjunto de sujeitos. Garantir o anonimato é impedir que um atacante seja capaz de identificar o sujeito quando seus dados forem associados com outras bases de dados e, associa-los com suas ações, caracterizando-se a anonimização dos dados.

A anonimização dos dados deve atender as exigências para atingir a preservação da privacidade, para isso, utiliza-se da remoção dos atributos identificadores, generalização dos dados para um nível de menor detalhe e supressão dos registros que não atendem ao k-anonimato, de forma a gerir os dados do sujeito e suas ações, preservando a privacidade e garantindo o anonimato.

Portanto, a técnica tem o objetivo de impedir o relacionamento entre os dados disponibilizados em base de dados pública com determinadas informações da vida do sujeito. Entretanto, este processo em muitos casos pode provocar perdas de dados, assim, acaba sendo um desafio disponibilizar, anonimizar e manter a utilidade dos dados.

A remoção dos atributos classificados como identificadores muitas vezes não é o suficiente para garantir o anonimato, pois hoje os atacantes podem utilizar de diversos meios para descobrir o sujeito. Como exemplificado neste trabalho, com o conhecimento prévio sobre o sujeito, o atacante pode combinar os atributos semi-identificadores e correlacionar com os atributos sensíveis e identificar o indivíduo.

A quantidade de semi-identificadores em uma tabela de dados também reflete na quebra de privacidade do sujeito, quanto mais semi-identificadores a tabela de dados possuir maior é a probabilidade de descaracterizar o anonimato.

Este trabalho apresentou exemplo de aplicação de processo de anonimização em uma base de dados, descrevendo suas principais características e indicando fragilidades e aspectos que devem ser considerados no processo de tornar acessíveis as bases de dados. Em função da importância e da inevitabilidade do acesso e compartilhamento dos dados, ganha relevância a questão do estudo, análise e divulgação dos fatores envolvidos com a alta disponibilização de dados e a privacidade.

REFERÊNCIAS

ANS. **Agência Nacional de Saúde Suplementar**. Disponível em: http://www.ans.gov.br/. Acesso em: 15 fev. de 2015.

BETTINI, Claudio; RIBONI, Daniele. **Privacy protection in pervasive system. State of art technical challenges**. Pervasive and Mobile Computing, 2014.

BORKO, Harold. **Information Science: What is it?** American Documentation, Jan, 1968.

BRANCO, Eliseu C Jr.; MACHADO, Javam. C.; MONTEIRO, José Maria. **Estratégicas para Proteção da Privacidade de Dados Armazenados na Nuvem**. Simpósio Brasileiro de Banco de dados. Tópicos em Gerenciamento de Dados e Informações 2014. Editora: Sociedade Brasileira de Computação (SBC), 1ª Ed. Capítulo 2. 2014.

BURKELL, Jacquelyn. **Anonymity in Behavioural Research: Not Being Unnamed, but Being Unknown**. University of Ottawa Law & Technology Journal, Vol. 3, No. 1, 2006.

CAMENISCH, Jan; FISCHER-HÜBNER, Simone; RANNENBERG, Kai. **Privacy and identity management for life**. Springer, 2011.

CAPURRO, Rafael; HJORLAND, Birger. The Concept of Information. Theorizing Information and Information Use. **Annual Review of informatiocn Science and Technology**. v. 37, cap. 8, p.343-411, 2003.

CAO, Jianneng. KARRAS, Panagiotis. **Publishing microdata with a robust privacy guarantee**. Proc. VLDB Endow, 5(11):1388–1399, 2012.

CHRISTOPHERSON, Kimberly M. Christopherson. **The positive and negative implications of anonymity in Internet social interactions: "On the Internet, Nobody Kobody Knows You´re a Dog.** Computers in Human Behavior 23, 2007.

CIRIANI, Valentina; VIMERCATI, Sabrina de Capitani Di; FORESTI, Sara; SAMARATI, Pierangela. **K-Anonymity**. Springer US, Advances in Information Security, 2007.

FOUCAULT, Michel. **Vigiar e punir**: nascimento da prisão. Tradução de Raquel Ramalhete. Petrópolis: Vozes, 1987.

FRIEDMAN, Arik; WOLFF, Ran.; SCHUSTER, Assaf. **Providing k-anonymity in data mining.** Published in: The VLDB Journal — The International Journal on Very Large Data Bases Volume 17 Issue 4, July 2008.

FUNG, Benjamin C.M.;WANG, Ke. FU, Ada Wai-Chee; YU, Philip S. **Introduction to Privacy-Preserving Data Publishing. Concepts and Techniques**. Chapman &Hall/CRC – Data Mining and Knowledge Discovery Series, 2010.

GÖRLICH, Werner A. **Big Data – Big Money, Big Risk – Big Tudo**. Mktcognitivo.com. 2015. Disponível em: <http://mktcognitivo.com/2015/02/06/big-data-big-money-big-risks-big-tudo>. Acesso em: 03 mar. 2015.

JENNINGS, Charles. Priv@cidade.com. **Como preservar sua intimidade na era da Internet**. São Paulo, 2000.

KAMBOURAKIS, Georgios. Anonymity and a closely related terms in the cyberspace: Na analysis by example. **Journal of Information Security and Applications**, n. I9 p. 2-17. 2014.

KORTH, Henry F. SILBERSCHATZ, Abraham. **Sistemas de Banco de Dados**. Makron Books, 1993.

LI, Ninghui; LI, Tiancheng; VENKATASUBRAMANIAN, Suresh. **t-closeness: Privacy beyond k-anonymity and ℓ-diversity**, in Proc. ICDE 2007, Istanbul, Turkey, 2007.

MACHANAVAJJHALA, Ashwin; GEHRKE, Johannes; KIFER, Daniel. **ℓ-Diversity: Privacy Beyond k-Anonymity**. Published in Journal ACM Transactions on Knowledge Discovery from Data (TKDD). Volume 1 Issue 1, Article No. 3, March 2007.

MOHAMMED, Noman; FUNG, Benjamin C. M; DEBBABI,Mourad. **Preserving Privacy and Utility in RFID Data Publishing**. 2010. Disponível em: http://spectrum.library.concordia.ca/6850/. Acesso em: Fev 2015.

NERGIZ,Mehmet Ercan; GÖK, Muhammed Zahit. **Hybrid K-Anonymity**. Computers & Security 44 (2014) 51-63. Elsevier, 2014.

PACHECO, Vinicius Maia. **Emprego de Anonimato para melhoria de privacidade no consumo de serviços em SAAS**. Tese de Doutorado – Universidade de Brasília, Faculdade de Tecnologia. Departamento de Engenharia Elétrica. 2013.

PFITZMANN, Andreas; KOHNTOPP, Marit. **Anonymity, unobservability, and pseudonymity—a proposal for terminology**. In Designing privacy enhancing technologies, pages 1–9. Springer, 2001.

RUN, Cui; KIM, Hyoung Joong; LEE, Dal-Ho; KIM ,Cheong Ghil; KIM, Kuinam J. **Protecting Privacy Using K-anonymity with a Hybrid Search Scheme**. Internacional Journal of Computer and Communication Engineering, Vol,1, nº 2, July 2012. Disponível em: http://www.ijcce.org/papers/41-Z022.pdf, 2012. Acesso em: 10 jan de 2015.

SAMARATI, Pierangela. **Protecting Respondents ´ Identities in Microdata Release**. IEEE Transactions on Knowledge and Data Engineering, Vol. 13, nº 6, 2001.

SAMARATI, Pierangela; SWEENEY, Latanya. **Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression**. From Technical Report SRI-CSL-98-04. Computer Science Laboratory, SRI. International, 1998.

SANT'ANA, Ricardo César Gonçalves. **Ciclo de Vida dos Dados e o papel da Ciência da Informação. In: XIV Encontro Nacional de Pesquisa em Ciência da Informação**, 2013, Florianópolis / SC. Anais do XIV Encontro Nacional de Pesquisa em Ciência da Informação, 2013. Disponível em: <http://enancib.sites.ufsc.br/index.php/enancib2013/XIVenancib/paper/viewFile/284/319> Acesso em: 20 dez. 2014.

SKOPEK, Jeffrey M. **Anonymity, the Production of Goods, and Institutional Design.** 82 Fordham L. Rev. 1751, 2014. Available at: http://ir.lawnet.fordham.edu/flr/vol82/iss4/4. Disponível em: <http://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=4960&context=flr> Acesso: 15 dez. 2014.

SWEENEY, Latanya. **k-anonymity: a model for protecting privacy**. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 557-570. 2002.

VIMERCATI, Sabrina de Capitani. FORESTI, Sara; LIVRAGA, Giovanni.; SAMARATI, Pierangela. **Data Privacy: Definicitions and Techiniques**. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems Vol. 20, No. 6 (2012) 793–817 World Scientific Publishing Company, 2012.

WEISE, Elizabeth. **Millions of Anthem customers alerted to hack**. USATODAY. Disponível em: <http://www.usatoday.com/story/tech/2015/02/05/anthem-health-care-computer-security-breach/22917635/>. Acesso em: 05 fev. de 2015.

WONG, Raymond Chi-Wing; LI, Jiuyong; FU, Ada Wai-Chee; WANG, Ke. **(α, k)-Anonymity: An Enhanced K-Anonymity Model for Privacy-Preserving Data Publishing**. KDD'0*6*, August 20–23, 2006, Philadelphia, Pennsylvania, USA, 2006.