Privacidade e Transparência no Setor Público: Um Estudo de Caso da Publicação de Microdados do INEP

Maria Jane de Queiroz[1], Gustavo H. M. B. Motta[1]

[1]Programa de Pós-Graduação em Informática – Universidade Federal da Paraíba (UFPB)
Caixa Postal 5115 – 58.051-970 – João Pessoa – PB – Brasil

jane.queiroz@ifrn.edu.br, gustavo@ci.ufpb.br

Abstract. *The Brazilian Freedom of Information Act determines the opening of data by the government, respecting the privacy of citizens. This case study analyzes an open government database, showing that unsystematic and ineffective forms of anonymization were used. Subsequently, systematic forms of anonymization are applied and a new analysis is done, showing the effectiveness of the procedures used. We conclude that an increasing attention by the government with regard to personal data is needed.*

Resumo. *A Lei de Acesso à Informação determina a abertura de dados pelo governo brasileiro, respeitando-se a privacidade dos cidadãos. Este estudo de caso analisa uma base de dados abertos governamental, mostrando que foram utilizadas formas assistemáticas e ineficazes de anonimização. Posteriormente, são aplicadas formas sistemáticas de anonimização e uma nova análise é efetuada, mostrando a eficácia dos procedimentos utilizados. Conclui-se que é necessária uma atenção maior do governo com relação aos dados pessoais.*

1. Introdução

O Brasil busca a transparência governamental desde 2004, quando a Controladoria-Geral da União criou portais de transparência. Em 2011, com a adesão do governo ao movimento *Open Government Data* (ou Dados Governamentais Abertos), sancionou-se a Lei nº 12.527 (ou Lei de Acesso à Informação – LAI) [Brasil 2011], definindo os moldes para a abertura de dados no Brasil [OGP 2013]. Segundo a LAI, a transparência é a regra e o sigilo, uma exceção. Entretanto, a LAI considera de forma especial os dados pessoais, conforme observado em seu Artigo 31, seção V, o qual afirma que o tratamento de dados pessoais deve respeitar a "intimidade, vida privada, honra e imagem das pessoas, bem como às liberdades e garantias individuais". Desse modo, a divulgação de dados pelo setor público não deve violar a privacidade dos cidadãos. Para que isso seja possível, faz-se necessária a aplicação de formas sistemáticas de anonimização, que dificultem a identificação individual em bases de dados disponíveis na Internet.

Este trabalho apresenta um estudo de caso visando verificar se a forma de anonimização adotada no setor público usa formas sistemáticas de anonimização. Nesse sentido, foram analisados os microdados preliminares do Censo da Educação Superior brasileira referente ao ano de 2013, realizado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira [INEP 2013]. Constatou-se que a anonimização dessa base é frágil e em seguida mostrou-se que a aplicação de formas sistemáticas de anonimização reduz eficazmente o risco de re-identificação de indivíduos. Espera-se com isso contribuir para o debate sobre o tratamento de dados pessoais por órgãos públicos. Ao final do artigo, são apresentadas as conclusões e os trabalhos futuros.

2. Formas de Anonimização

Anonimizar significa ocultar a identidade e/ou dados confidenciais das pessoas em bases de dados, sem grandes prejuízos à utilidade da informação [Fung et al. 2011]. Existem duas formas de anonimização: assistemática e sistemática. A primeira consiste na mera exclusão de dados explícitos, que identificam uma pessoa de forma individual (como nome completo, CPF e outras informações pessoais). Essa forma de anonimização não assegura a preservação da privacidade do indivíduo.

Já a forma sistemática, empregada neste estudo de caso, utiliza uma série de técnicas e modelos de anonimização eficazes, que dificultam a re-identificação individual a partir de uma base de dados publicada na Internet. Para isso, são realizados os seguintes procedimentos: classificação de atributos, definição de operações, aplicação de modelos de anonimização, análise dos resultados quanto à preservação de privacidade e à utilidade dos dados anonimizados. Neste estudo de caso, são aplicados os modelos de anonimização *k-Anonymity* e *Distinct ℓ-Diversity*. O primeiro é aplicado aos atributos com potencial para levar à re-identificação (QIDs ou *Quasi Identifiers*) e cria grupos com *k* registros idênticos quanto aos QIDs. O segundo é aplicado a atributos confidenciais e define que em cada grupo deve haver ℓ valores distintos quanto a estes atributos. Informações adicionais podem ser encontradas em Fung et al. (2011).

3. Resultados

Esta seção apresenta os resultados obtidos a partir da análise das fragilidades da base de dados do Censo da Educação Superior brasileira de 2013 [INEP 2013], seguida da aplicação de técnicas sistemáticas de anonimização a tal base.

3.1. Análise da base de dados original

Para verificar a forma de tratamento dispensada a dados pessoais por órgãos públicos, foram utilizados os microdados do Censo da Educação Superior brasileira de 2013, cujos dados foram coletados, tratados e disponibilizados pelo INEP. Dentre os arquivos disponibilizados em formato CSV (*Comma Separated Values*) com dados sobre docentes, alunos e instituições de ensino superior (IESs), foi selecionado para análise o arquivo DM_DOCENTE, contendo 49 colunas e 383.683 linhas com informações pessoais, profissionais e acadêmicas dos docentes.

Tal arquivo foi importado para o programa *Microsoft Access* e os dados de um dos autores deste artigo, que atua como docente em uma IES, foram filtrados utilizando-se a opção "Filtrar por formulário". Os valores aplicados ao filtro foram {17/12/1988; feminino; Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte}, correspondentes à data de nascimento, sexo, nome da IES. Após a aplicação do filtro, foi retornado apenas um registro, possibilitando a re-identificação individual e a consequente violação da privacidade da docente, visto que a base de dados contém informações adicionais, como idade completa (neste caso, 25 anos) e nível de escolaridade (neste caso, especialização) e informações sensíveis – indicando, por exemplo, se um docente possui deficiência e, em caso afirmativo, o tipo de deficiência.

Com esta análise foi possível mostrar que a base de dados disponibilizada pelo INEP utiliza formas assistemáticas de anonimização e que tais formas são ineficazes, pois apesar da supressão do nome completo dos docentes (e possivelmente, de outras informações sensíveis), ainda é possível a re-identificação individual.

3.2. Anonimização da base de dados original

A partir da verificação da vulnerabilidade da base de dados original, exposta na seção anterior, a planilha DM_DOCENTE foi submetida a um processo de anonimização sistemática, sendo aplicados os procedimentos citados na seção 2, como classificação de atributos, definição de operações (generalização e supressão de dados) e aplicação de modelos clássicos de anonimização (*k-Anonymity* e *Distinct ℓ-Diversity*).

Esses procedimentos foram aplicados com o auxílio da ferramenta ARX, versão 3.2.1, desenvolvida por Prasser et al. (2015). Para tanto, a ferramenta foi configurada para utilizar o valor máximo e mínimo possíveis para k (ou seja, $k = 2$ e $k = 100$) e ℓ (ou seja, $\ell = 2$ e $\ell = 9$). Os resultados obtidos após a transformação anônima são apresentados na Tabela 1.

A propriedade perda de informação é calculada de acordo com a métrica *Discernibility*, que cobra uma penalidade por cada registro idêntico quanto aos valores de QID [Fung et al. 2011]. Na Tabela 1, é possível observar que quanto maior a quantidade de tuplas suprimidas, maior a perda de informação na base de dados anônima. A propriedade classes de equivalência compreende os grupos criados com QIDs idênticos a partir da execução dos dois modelos (*k-Anonymity* e *Distinct ℓ-Diversity*) utilizados em conjunto, indicando quantos grupos foram criados em cada transformação. A propriedade classes suprimidas indica a quantidade de grupos que tiveram seus registros (também chamados de linhas ou tuplas) suprimidos.

Tabela 1. Transformações anônimas resultantes.

Propriedade	***k*=ℓ = 2**	***k*= 100, ℓ = 2**	***k*= 100, ℓ = 9**
Perda de informação	1.0138272263E10	3.554405535E9	5.1157096261E10
Tuplas suprimidas	23234	4445	24260
Classes de equivalência	2638	337	5
Classes suprimidas	1757	122	1
Tamanho mínimo da classe	2	101	34348
Tamanho máximo da classe	13333	12883	168376

Com relação aos riscos de re-identificação, a ferramenta ARX utiliza os tamanhos das classes de equivalência para estimar as probabilidades de re-identificação. Os valores para os riscos de re-identificação obtidos para as transformações apresentadas anteriormente são mostrados na Tabela 2.

Tabela 2. Riscos de re-identificação individual em cada transformação.

		Antes da Anonimização		*Depois da anonimização*	
Valor de k	***Valor de ℓ***	***Alto***	***Baixo***	***Alto***	***Baixo***
2	*2*			50%	0,00430%
100	*2*	100%	33%	0,99010%	0,00776%
100	*9*			0,00412%	0,00059%

Observa-se que quanto maior o valor de k e ℓ, menores os valores para os riscos mais altos e mais baixos de re-identificação individual, com o ônus do aumento da perda de informação decorrente de supressões e generalizações na base de dados anônima.

3.3. Análise da base de dados após a anonimização

A base de dados anônima foi importada para o *software Microsoft Access*, a fim de realizar uma nova análise para verificar se os procedimentos utilizados foram eficazes. Para tanto, foi utilizada a transformação com os valores de $k = \ell = 2$, por apresentar o maior valor para alto risco de re-identificação individual e consequentemente, uma alta probabilidade de sucesso na re-identificação individual.

Novamente foi utilizada a opção "Filtrar por formulário" do programa *Microsoft Access*, porém como os valores para nome da IES, dia e mês de nascimento foram suprimidos, essas informações não puderam mais ser utilizadas. O ano de nascimento passou a ser representado em forma de intervalos, bem como a idade dos docentes.

Assim, nesta nova análise, filtraram-se os valores {Nordeste, Feminino, [1985-1990[, [25-32[}, equivalentes respectivamente à região onde se situa a IES, sexo, ano e idade da docente procurada. Como a análise usa dados reais e, na primeira análise, a própria base de dados continha a idade completa da docente procurada (25 anos), optou-se por utilizar tal informação na nova filtragem. Como resultados, foram obtidos 2764 registros, tornando inexequível a re-identificação individual da docente e mostrando que soluções de anonimização sistemáticas podem ser eficazes em casos reais. De fato, após a anonimização da base, o risco mais alto de re-identificação, para $k = \ell = 2$, é de 50% e o mais baixo é 0,00430% (Tabela 2).

4. Conclusão e trabalhos futuros

Os resultados deste estudo de caso destacam a necessidade de uma maior atenção, por parte dos órgãos da administração pública, para a preservação de privacidade em dados pessoais dos cidadãos quando da abertura de dados. Uma vez exposto o problema e as soluções existentes, são necessários estudos aprofundados para definir procedimentos padrão de anonimização, viáveis de serem aplicados nas bases de dados do governo, bem como para elaborar políticas de capacitação dos servidores responsáveis pela coleta e publicação de dados, para que se adotem medidas de preservação de privacidade nesse processo.

Referências

Brasil. (2011). "Lei nº 12.527, de 18 de novembro de 2011". http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm

Fung, B. C. M., Wang, K., Fu, A. W. C., Yu, P. S. (2011). "Introduction to Privacy-Preserving Data Publishing", Chapman & Hall/CRC.

INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira). (2013). "Microdados para download", http://portal.inep.gov.br/basica-levantamentos-acessar

OGP (*Open Government Partnership*). (2013). "2º Plano de Ação Brasileiro", http://edemocracia.camara.gov.br/documents/980199/980230/2%C2%BA%20Plano+de+A%C3%A7%C3%A3o/.

Prasser, F., Kohlmayer, F., Babioch, K., Xhani, L., Dshevlekov, L., Schneider, M. (2015). "ARX - Data Anonymization Tool", http://arx.deidentifier.org/.

