Análise de Riscos à Privacidade de Indivíduos Presentes em Bases de Dados Estatísticas do IBGE

Lucas Caetano Lopes Rodrigues[1]

[1]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG)

Resumo. *Para garantir que divulgações estatísticas realizadas por órgãos governamentais estão de acordo com a Lei Geral de Proteção de Dados e que os dados dos indivíduos coletados e divulgados pelas instituições não poderão ser utilizados de maneira indevida, é necessária a análise cuidadosa dos riscos à privacidade individual incorridos pela divulgação dessas pesquisas. Neste trabalho, criamos um modelo de ataque de membership, no qual um adversário tenta identificar a presença ou ausência de um indivíduo em uma amostra, implementamos e executamos esse ataque sobre a base de dados da Pesquisa Nacional de Saúde do Escolar (PeNSE) divulgada pelo Instituto Brasileiro de Geografia e Estatística em 2015, revelando que indivíduos presentes na base de dados do PeNSE estão vulneráveis ao ataque de membership, e a amostragem não é uma técnica rigorosa para garantia de privacidade de todos os indivíduos em uma base de dados.*

1. Introdução

Com o intuito de prover informações relevantes para pesquisadores e tomadores de decisões no poder público, órgãos governamentais coletam e divulgam estatísticas sobre a população do país em diversas esferas da sociedade. Dois dos principais órgãos que desempenham esse trabalho no Brasil são o Instituto Brasileiro de Geografia e Estatística (IBGE), que realiza pesquisas estatísticas em diversos âmbitos, como pesquisas domiciliares e o Censo Demográfico, e o e o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), que realiza pesquisas estatísticas na área da Educação no Brasil.

Para realizar essas pesquisas, os órgãos governamentais coletam uma grande quantidade de dados da população através de formulários digitais, questionários físicos, e visitas a domicílios. Os dados coletados frequentemente possuem informações sensíveis sobre indivíduos, como condições de saúde, situação financeira da família, uso de drogas ilícitas, etc. Indivíduos que cedem seus dados para essas pesquisas estão sujeitos à possível utilização indevida dos dados sensíveis, e podem se sentir inclinados a oferecerem informações falsas, diminuindo a utilidade estatística das pesquisas realizadas pelos institutos. Neste sentido, torna-se necessária a garantia de proteção à privacidade dos dados individuais como forma de incentivo aos participantes da pesquisa para que forneçam informações legítimas, ao mesmo tempo protegendo os indivíduos contra o uso indevido de seus dados pessoais.

Este projeto visa investigar os riscos à privacidade aos quais indivíduos que participam das pesquisas estatísticas realizadas e divulgadas pelo IBGE estão submetidos, em particular a Pesquisa Nacional de Saúde do Escolar (PeNSE). Para isso, definimos o modelo de ataque de *membership*, que utiliza a base de dados do Censo Educacional do Inep

para decidir a presença ou ausência de indivíduos nas bases de dados amostrais do PeNSE, e propomos duas medidas de degradação de privacidade de indivíduos presentes em bases de dados amostrais que podem ser aplicadas nas bases do PeNSE, a *degradação de privacidade* e a *degradação de privacidade esperada*. Por fim, implementamos e executamos os ataques sobre a base de dados do PeNSE e do Censo Escolar de 2015, revelando que, apesar de proteger a privacidade de grande parte dos indivíduos presentes na base, a amostragem como técnica de proteção à privacidade não consegue impedir um adversário de descobrir informações sensíveis sobre alguns indivíduos.

O restante desse trabalho está organizado da seguinte forma: a Seção 2 possui o referencial teórico coletado da literatura sobre controle de divulgação estatística. A Seção 3 contempla as contribuições desse projeto de pesquisa, a saber, as premissas assumidas para o ataque, a descrição do modelo de adversário e do ataque que ele conduz sobre as bases de dados, as medida de degradação de privacidade, e os resultados encontrados ao executar o ataque sobre a base de dados do PeNSE e do Censo Escolar de 2015. A Seção 4 conclui o trabalho, indicando direções para futuras pesquisas no campo de controle de divulgação estatística para bases de dados amostrais.

2. Trabalhos Relacionados e Referencial Teórico

O campo de estudo de privacidade em bases de dados estatísticas tem visto bastante crescimento em interesse na última década, e um grande esforço tem sido direcionado à modelagem e desenvolvimento de técnicas para avaliação de riscos à privacidade e métodos para mitigar esses riscos, tanto na academia quanto na indústria.

A seguir, apresentamos os conceitos e definições pertinentes para o projeto presentes na literatura técnica [Dwork 2011] [Alvim et al. 2019] [Ghosh et al. 2009] [Kifer and Machanavajjhala 2011].

- **Base de dados estatística** é uma base de dados que contém informações agregadas de indivíduos de uma população em forma de microdados ou dados tabulares. Em geral, bases de dados estatísticas são utilizadas para tomadas de decisões informadas pelo poder público e para pesquisas acadêmicas sobre a população. No contexto deste Projeto de Pesquisa, trabalhamos com bases de dados estatísticas divulgadas por órgãos públicos, em particular o IBGE e o Inep, de livre acesso para a população.
- **Base de dados amostral** é uma base de dados resultante do pós-processamento de uma base de dados estatística que contém uma fração dos registros da base de dados original. As técnicas utilizadas para gerar a base de dados amostral não são relevantes no contexto deste Projeto de Pesquisa.
- **Atributo sensível** é um atributo ou conjunto de atributos na base de dados que revela informações sensíveis sobre os indivíduos presentes na base. Os atributos considerados sensíveis podem mudar dependendo do tipo de base de dados, do objetivo do adversário, e dos interesses dos indivíduos presentes na base. Exemplos de atributo sensíveis são: condições de saúde, renda mensal do indivíduo ou da família do indivíduo, quantidade de eletrodomésticos na residência do indivíduo, e frequência de uso de drogas ilícitas.
- **Quaseidentificadores** são conjuntos de atributos na base de dados que possibilitam a associação de um registro presente na base de dados a um indivíduo através de um ataque de reidentificação.

- **Informação lateral** é qualquer conjunto de informações disponíveis publicamente, fora da base de dados de interesse, sobre indivíduos, e que auxiliem em um ataque de reidentificação.
- **Indivíduo-alvo** é um indivíduo ou conjunto de indivíduos que um adversário pode ser capaz de reidentificar através de um ataque de reidentificação intencional ou não intencional.
- **Adversário** é uma pessoa ou grupo de pessoas que possui acesso à base de dados estatística de interesse e à informação lateral sobre um indivíduo-alvo, e pode utilizar essas informações em um ataque de reidentificação para associar registros presentes na base de dados a indivíduos. Um adversário não necessariamente possui intenções malignas, mas neste Projeto de Pesquisa consideramos que a privacidade dos indivíduos presentes nas bases de dados estatísticas deve ser protegida independentemente das intenções do adversário.
- **Ataque de membership** é um ataque realizado pelo adversário cujo objetivo é identificar se o indivíduo-alvo está ou não presente na base de dados amostral gerada a partir da base de dados estatística.
- **Ataque de reidentificação** é um ataque realizado pelo adversário sobre uma base de dados estatística que cruza os dados providos pela base e os dados advindos de outros meios (i.e., informação lateral) para associar registros presentes na base de dados com indivíduos.
- **Ataque de inferência de atributo** é um ataque realizado pelo adversário sobre uma base de dados estatística que cruza os dados providos pela base e os dados advindos de outros meios (i.e., informação lateral) para inferir atributos sensíveis de indivíduos-alvo.
- **Riscos à privacidade individual** são os riscos que incorrem da publicação da base de dados estatística sobre um indivíduo-alvo. Informalmente, uma medida de risco à privacidade individual é uma comparação entre a probabilidade de reidentificação do indivíduo na base de dados *antes* e *depois* de cruzar as informações disponíveis na base com as informações laterais adquiridas através de outros meios.
- **Conhecimento a priori** é o conhecimento que o adversário possui sobre o indivíduo-alvo antes da execução do ataque. No contexto desse Projeto de Pesquisa, assumimos que o adversário possui como conhecimento a priori a base de dados estatística completa.
- **Conhecimento a posteriori** é o conhecimento que o adversário possui sobre o indivíduo-alvo após a execução do ataque. No contexto desse Projeto de Pesquisa, o adversário utiliza seu conhecimento a priori e a informação lateral em conjunto com a divulgação da base de dados amostral para adquirir informações sensíveis sobre o indivíduo-alvo.
- **Degradação de privacidade** é uma medida do risco que incorre da publicação da base de dados amostral sobre um indivíduo-alvo. Informalmente, uma medida de risco à privacidade individual é uma comparação entre a probabilidade de reidentificação do indivíduo na base de dados estatística *antes* e *depois* da publicação da base de dados amostral. Após a publicação, o adversário pode utilizar os dados contidos na base de dados amostral para realizar um ataque de reidentificação e descobrir atributos sensíveis sobre o indivíduo-alvo. A Figura 1 mostra como a degradação de privacidade se relaciona com a publicação da base

de dados amostral.

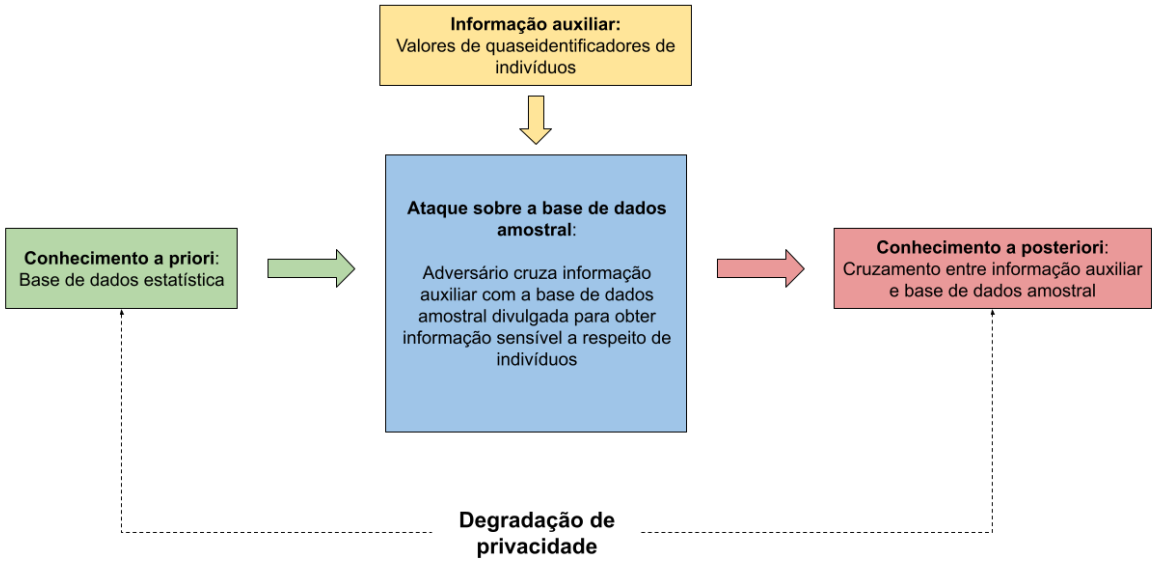


Figura 1. Degradação de privacidade causada pela divulgação da base de dados amostral.

3. Contribuições

Esta seção apresenta as contribuições deste Projeto de Pesquisa, a saber, a descrição do modelo de ataque empregado pelo adversário para identificar se um indivíduo-alvo está ou não presente na base de dados amostral, as medida de risco à privacidade individual e coletiva causados pela divulgação da base de dados amostral, e a descrição dos experimentos realizados ao executar o ataque sobre as bases de dados do Censo Escolar e da Pesquisa Nacional de Saúde do Escolar.

Iniciamos a discussão apresentando o cenário completo da divulgação da base de dados amostral. Em seguida, descrevemos o modelo do ataque de *membership*, incluindo as premissas utilizadas e o conhecimento prévio que o adversário precisa para realizar o ataque, e calculamos o risco à privacidade individual incorrido pela divulgação da base de dados amostral utilizando um exemplo condutor apresentado ao longo da seção. Por fim, descreveremos como o ataque foi realizado sobre as bases de dados do Censo Escolar e da Pesquisa Nacional de Saúde do Escolar e apresentamos uma discussão sobre os resultados encontrados.

3.1. Divulgação da base de dados amostral

Consideramos uma base de dados amostral, S, extraída de uma base de dados estatística através de uma amostragem uniforme. Ou seja, se a base de dados estatística, D, possui $|D|$ registros, cada um deles pode ser selecionado para a amostra com probabilidade $1/|D|$.

Sobre a base de dados estatística D, assumimos que:

- **Premissa 1: Todo indivíduo presente na base de dados estatística possui apenas um único registro associado a ele**. Ou seja, nenhum indivíduo está presente na base de dados mais de uma vez.

- **Premissa 2: A base de dados estatística possui todos os indivíduos de interesse no contexto do ataque**. Por exemplo, para um adversário que possui como indivíduo-alvo um professor do Departamento de Ciência da Computação da UFMG, ainda que esse professor não esteja na base de dados amostral, consideramos que ele certamente está na base de dados estatística da qual a amostra foi gerada. No contexto desse projeto, essa premissa é razoável pois as bases de dados do Inep publicadas anualmente possuem todos os indivíduos (professores e alunos) matriculados em instituições de ensino do Brasil. Além disso, as bases de dados amostrais com a qual trabalhamos, publicada pelo IBGE na Pesquisa Nacional de Saúde do Escolar, possui um subconjunto dos indivíduos matriculados em instituições de ensino do Brasil.

Consideramos também que a base de dados amostral possui um subconjunto dos atributos da base de dados estatística, ao qual são adicionados outros atributos, possivelmente sensíveis, e que o tamanho da amostra, $|S|$, é fixo. A Tabela 1 apresenta um exemplo de base de dados estatística D que contém todos os indivíduos de uma população de interesse e a Tabela 2 apresenta uma base de dados amostral de tamanho 4 dos indivíduos em D.

ID	ESTADO	SEXO	IDADE	ATRIBUTO SENSÍVEL 1
1	SP	F	31-40	0
2	MG	F	21-30	1
3	MG	F	21-30	1
4	RJ	M	21-30	0
5	RJ	M	21-30	1
6	RJ	M	21-30	0
7	SP	M	31-40	0
8	SP	M	31-40	1
9	SP	M	31-40	0
10	SP	M	31-40	1

Tabela 1. Base de dados estatística da população de interesse, D.

ID	ESTADO	SEXO	IDADE	ATRIBUTO SENSÍVEL 2
A	SP	F	31-40	1
B	MG	F	21-30	0
C	RJ	M	21-30	0
D	RJ	M	21-30	1

Tabela 2. Base de dados amostral S construída através da amostragem uniforme dos registros da base de dados estatística subjacente, D, da qual é removida a coluna de Atributo Sensível 1, e adicionada a coluna de Atributo Sensível 2, dado o contexto da publicação da base de dados amostral.

3.2. Adversário

Para descrever o comportamento do adversário, precisamos definir o seu conhecimento a priori, a informação lateral que ele possui e como o ataque é conduzido. Nesta seção

apresentamos as premissas assumidas sobre o adversário e nas seções seguintes apresentamos o modelo do ataque e o cálculo dos riscos à privacidade incorridos pela divulgação da base de dados amostral.

Sobre o adversário, assumimos que:

- **Premissa 3: O adversário conhece a base de dados estatística** D. No contexto desse projeto, essa premissa indica que o adversário tem acesso à base de dados do Censo Escolar publicada pelo Inep.
- **Premissa 4: O adversário conhece um conjunto de quaseidentificadores para o indivíduo-alvo**. O adversário tem o conhecimento de um conjunto de atributos da base de dados estatística que pode reidentificar o indivíduo-alvo (i.e., quaseidentificadores). Como não sabemos a priori qual o conjunto de quaseidentificadores o adversário possui, não colocamos restrições sobre esse conjunto. Note que esse conjunto de quaseidentificadores não necessariamente identifica unicamente o indivíduo-alvo na base de dados estatística.

A base de dados estatística D pode possuir atributos sensíveis sobre indivíduos. Entretanto, no contexto desse projeto, assumimos que o adversário utiliza a base de dados estatística apenas para realizar um ataque à base de dados amostral, possivelmente para descobrir outros atributos sensíveis contidas nesta.

A Figura 2 mostra, no exemplo condutor, um exemplo do conhecimento prévio e da informação lateral que o adversário possui: a base de dados estatística D e os seguintes quaseidentificadores para o indivíduo-alvo $\{\texttt{ESTADO} = \texttt{SP}, \texttt{SEXO} = \texttt{F}, \texttt{IDADE} = 31 - 40\}$. Através desses quaseidentificadores, o adversário sabe que o indivíduo-alvo possui $\texttt{ID} = 1$ na base de dados estatística, e o valor do Atributo Sensível 1 é 0. Seu objetivo é, a princípio, identificar se o indivíduo-alvo está na base de dados amostral S, e em seguida descobrir o valor do atributo sensível 2.

ID	ESTADO	SEXO	IDADE	ATRIBUTO SENSÍVEL 1
1	SP	F	31-40	0
2	MG	F	21-30	1
3	MG	F	21-30	1
4	RJ	M	21-30	0
5	RJ	M	21-30	1
6	RJ	M	21-30	0
7	SP	M	31-40	0
8	SP	M	31-40	1
9	SP	M	31-40	0
10	SP	M	31-40	1

Figura 2. Conhecimento prévio do adversário. O adversário conhece a base de dados estatística D e os seguintes valores de quaseidentificadores para o indivíduo-alvo: $\{\texttt{ESTADO} = \texttt{SP}, \texttt{SEXO} = \texttt{F}, \texttt{IDADE} = 31 - 40\}$.

3.3. Ataques sobre bases de dados amostrais

Como apresentado na Seção 2, para realizar ataques sobre bases de dados estatísticas, o adversário frequentemente tenta **reidentificar o indivíduo-alvo** ou **inferir um atributo sensível do indivíduo-alvo**. Esses dois ataques, entretanto, necessitam de uma premissa que não podemos assumir quando estamos tratando de bases de dados amostrais: o registro associado ao indivíduo-alvo deve estar presente na base de dados alvo do ataque.

Portanto, para bases de dados amostrais, é interessante que o adversário comece seu ataque identificando a presença ou ausência do indivíduo-alvo, para então realizar os dois ataques citados no parágrafo anterior. Dito isso, neste projeto de pesquisa, o nosso foco está em construir este ataque inicial que identifica a presença ou ausência do indivíduo no base de dados amostral, chamado **ataque de *membership***. Além da justificativa anterior, ao longo do projeto percebemos também que, muitas vezes, o ataque de *membership* pode ser equivalente ao ataque de reidentificação quando, por exemplo, o indivíduo é único na base de dados estatística e está presente na base de dados amostral.

Em um ataque de *membership* o adversário tem acesso completo à base de dados estatística que contém todos os indivíduos de interesse. Além disso, o adversário possui conhecimento dos valores de um conjunto de quaseidentificadores do indivíduo-alvo. Seu objetivo é identificar se o indivíduo-alvo está ou não presente na base de dados amostral gerada a partir da base de dados estatística. A seguir, definimos como o adversário conduz o ataque de *membership*.

3.4. Ataque de *membership* individual

Para avaliar o risco à privacidade de um indivíduo causado pela divulgação da base de dados amostral, comparamos o sucesso do adversário ao conduzir o ataque antes e após a base de dados amostral ser divulgada. Para isso, definimos as seguintes métricas para o ataque de *membership*:

- **Sucesso a priori**: O sucesso que o adversário tem de supor corretamente a presença ou ausência do indivíduo na amostra, antes da divulgação da base de dados amostral.
- **Sucesso a posteriori**: O sucesso que o adversário tem de identificar corretamente a presença ou ausência do indivíduo na amostra, depois da divulgação da base de dados amostral.
- **Degradação de privacidade**: Uma comparação entre o sucesso a priori e o sucesso a posteriori do adversário.

3.4.1. Sucesso a priori

Antes da base de dados amostral ser divulgada, o adversário pode apenas supor que o indivíduo-alvo será incluído nela com uma certa probabilidade. Na ausência do conhecimento sobre a técnica de amostragem utilizada, o adversário assume que a amostragem será feita de maneira uniforme sobre os registros da base de dados estatística[1].

[1]Esta decisão por parte do adversário parte do Princípio da Máxima Entropia, que diz que a melhor distribuição de probabilidade para representar o conhecimento sobre um sistema é aquela com maior entropia.

Portanto, para medir o **sucesso a priori**, comparamos o número total de amostras possíveis da base de dados D, restringindo o tamanho da amostra a $|S|$, com o número total de amostras em que o indivíduo-alvo está presente.

O número de maneiras de escolher $|S|$ registros de uma base de dados com $|D|$ registros, que chamaremos de σ, é

$$\sigma = \binom{|D|}{|S|} = \frac{|D|!}{|S|!(|D| - |S|)!}$$

Como queremos comparar o número total de amostras possíveis com a quantidade destas amostras em que o indivíduo-alvo está presente, fixamos-no na amostra e contamos de quantas formas podemos selecionar os outros $|S| - 1$ indivíduos de D. Este valor chamamos de θ

$$\theta = \binom{|D| - 1}{|S| - 1} = \frac{(|D| - 1)!}{(|S| - 1)!(|D| - |S| - 1)!}$$

Por fim, a fração de amostras em que o indivíduo-alvo está presente é

$$\frac{\theta}{\sigma} = \frac{\binom{|D|-1}{|S|-1}}{\binom{|D|}{|S|}} = \frac{\frac{(|D|-1)!}{(|S|-1)!(|D|-|S|-1)!}}{\frac{|D|!}{|S|!(|D|-|S|)!}} = \frac{|S|}{|D|}$$

Definimos o **sucesso a priori** como esse valor, θ/σ.

$$\texttt{sucesso-priori} = \frac{|S|}{|D|}$$

No exemplo condutor, cuja base de dados D possui 10 registros, e a amostra S possui 4 registros, o sucesso a priori do adversário é $4/10 = 0.4$.

3.4.2. Sucesso a posteriori

Após a divulgação da base de dados amostral, o adversário conduz o ataque para tentar identificar a presença ou ausência do indivíduo-alvo na amostra. Com essa finalidade, o adversário realiza os seguintes passos durante o ataque:

1. **Seleciona os indivíduos na base de dados estatística utilizando os quaseidentificadores**. Nesta etapa, o adversário realiza uma seleção na base de dados estatística filtrando os valores dos atributos pelos quaseidentificadores que ele conhece. A base de dados estatística pode possuir 1 ou mais registros com os valores de atributos especificados. Chamaremos a quantidade de indivíduos selecionados neste passo de $n : 1 \leq n \leq |D|$.
2. **Seleciona os indivíduos na base de dados amostral utilizando os quaseidentificadores**. Nesta etapa, o adversário realiza uma seleção na base de dados amostral filtrando os valores dos atributos pelos quaseidentificadores que ele conhece. A

base de dados amostral pode possuir 0 ou mais registros com os valores de atributos especificados. Chamaremos a quantidade de indivíduos selecionados neste passo de $d : 0 \leq d \leq |S|$.

Após a execução do segundo passo, o adversário pode se encontrar nos seguintes estados:

- A seleção do passo 1 retorna apenas um indivíduo, e a seleção do passo 2 retorna 0 indivíduos. Nesse caso, $n = 1$ e $d = 0$.
- A seleção do passo 1 retorna apenas um indivíduo, e a seleção do passo 2 retorna apenas 1 indivíduo. Nesse caso, $n = 1$ e $d = 1$.
- A seleção do passo 1 retorna mais de um indivíduo, e a seleção do passo 2 retorna 0 indivíduos. Nesse caso, $1 < n \leq |D|$ e $d = 0$.
- A seleção do passo 1 retorna mais de um indivíduo, e a seleção do passo 2 retorna apenas 1 indivíduo. Nesse caso, $1 < n \leq |D|$ e $d = 1$.
- A seleção do passo 1 retorna mais de um indivíduo, e a seleção do passo 2 retorna mais de 1 indivíduo. Nesse caso, $1 < n \leq |D|$ e $1 < d \leq |S|$.

Para os casos em que $d = 0$, o adversário identifica corretamente que o indivíduo não foi selecionado para a amostra. Para o contexto desse projeto, o ataque foi bem sucedido, mas podemos dizer que não há vazamento de dados ou riscos à privacidade incorridos pela divulgação da base de dados amostral.

Para os casos em que $d \neq 0$, consideramos que o adversário pode supor que um dos registros encontrados na base de dados amostral é o indivíduo-alvo com probabilidade n/d. Definimos, portanto, o sucesso a posteriori como

$$\texttt{sucesso-posteriori} = \begin{cases} 0 & \text{se } d = 0 \\ n/d & \text{caso contrário} \end{cases}$$

A Figura 3 mostra como esse ataque seria realizado no exemplo condutor. A seleção realizada pelo primeiro passo do ataque retorna apenas um indivíduo na base de dados estatística. A seleção realizada pelo segundo passo retorna apenas um indivíduo na base de dados amostral. Portanto, o sucesso a posteriori é $n/d = 1/1 = 1$.

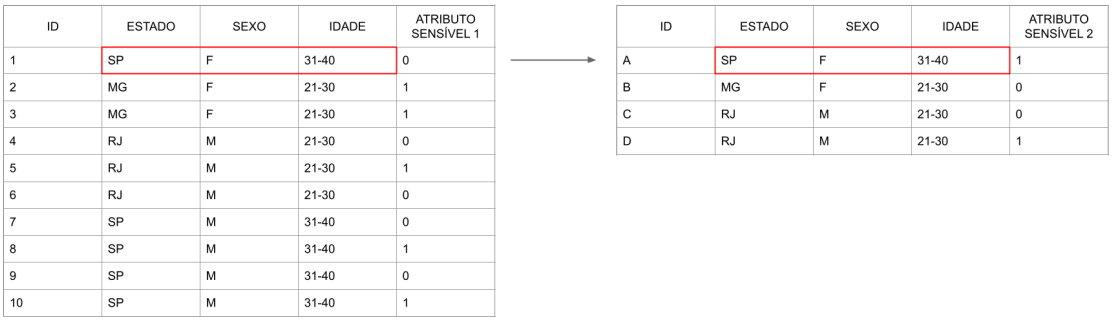
3.4.3. Degradação de privacidade

Por fim, definimos a **degradação de privacidade** como uma comparação do sucesso a priori com o sucesso a posteriori. A degradação de privacidade mede qual o risco à privacidade incorrido a um indivíduo-alvo pela publicação da base de dados amostral.

$$\texttt{degradação-privacidade} = \frac{\texttt{sucesso-posteriori}}{\texttt{sucesso-priori}}$$

No nosso exemplo condutor, a degradação de privacidade é

$$\texttt{degradação-privacidade} = \frac{\texttt{sucesso-posteriori}}{\texttt{sucesso-priori}} = \frac{1}{0.4} = 2.5$$



ID	ESTADO	SEXO	IDADE	ATRIBUTO SENSÍVEL 1
1	SP	F	31-40	0
2	MG	F	21-30	1
3	MG	F	21-30	1
4	RJ	M	21-30	0
5	RJ	M	21-30	1
6	RJ	M	21-30	0
7	SP	M	31-40	0
8	SP	M	31-40	1
9	SP	M	31-40	0
10	SP	M	31-40	1

ID	ESTADO	SEXO	IDADE	ATRIBUTO SENSÍVEL 2
A	SP	F	31-40	1
B	MG	F	21-30	0
C	RJ	M	21-30	0
D	RJ	M	21-30	1

Figura 3. Sucesso a posteriori do adversário, ao realizar o ataque descrito na Seção 3.4.2. Neste caso, o adversário encontra um único indivíduo para a seleção do passo 1, e um único indivíduo para a seleção do passo 2, i.e., $n = 1, d = 1$. O sucesso a posteriori é $n/d = 1/1 = 1$

.

Podemos dizer que a divulgação da base de dados amostral do exemplo aumentou o sucesso do adversário em identificar a presença do indivíduo-alvo na amostra em 2.5 vezes.

3.5. Ataque de *membership* coletivo

Como discutido na **Premissa 4** da Seção 3.2, neste projeto de pesquisa não assumimos nada sobre o conhecimento prévio do adversário sobre os quaseidentificadores do indivíduo-alvo. Portanto, nesta Seção, definimos o ataque de *membership* coletivo, que leva em consideração todos os conjuntos de quaseidentificadores possíveis para o indivíduo-alvo. Este ataque não se assemelha ao comportamento real de um adversário que deseja realizar um ataque sobre as bases de dados, entretanto ele tem grande valor para uma análise genérica dos riscos incorridos aos indivíduos em uma base de dados amostral.

Neste ataque, definimos todas as partições geradas por quaseidentificadores na base de dados estatística e na base de dados amostral, e contamos o número de indivíduos dentro dessas partições. Para cada conjunto de quaseidentificadores, calculamos qual é a degradação de privacidade para um indivíduo-alvo que possui esse conjunto de quaseidentificadores. Em seguida, somamos esses valores ponderados pela probabilidade de aquele indivíduo, presente na amostra, ser selecionado como indivíduo-alvo.

Desta forma, definimos uma nova medida de degradação de privacidade: a ***degradação de privacidade esperada***. Sejam $i = \{1, 2, ...\}$ as partições geradas por um conjunto de quaseidentificadores na base de dados estatística e na amostra. Para cada partição, temos n_i e d_i, a quantidade de indivíduos presentes na partição i da base de dados estatística e da amostra, respectivamente.

$$E[\texttt{degradação-privacidade}] = \sum_i \frac{1}{|S|} \cdot \texttt{degradação-privacidade}_i$$

$$E[\texttt{degradação-privacidade}] = \sum_i \frac{1}{|S|} \cdot \frac{\texttt{sucesso-posteriori}_i}{\texttt{sucesso-priori}}$$

$$\texttt{degradação-esperada} = E[\texttt{degradação-privacidade}] = \sum_i \frac{1}{|S|} \cdot \frac{n_i/d_i}{|S|/|D|}$$

As Figuras 4 e 5 mostram as partições geradas pelos quaseidentificadores de interesse nas bases de dados estatística e amostral do exemplo condutor, respectivamente. Para cada indivíduo na amostra, a probabilidade de ele ser selecionado como indivíduo-alvo é $1/|S| = 1/4$. Além disso, para cada uma das partições, calculamos seu tamanho na amostra, n_i, e seu tamanho na base de dados estatística, d_i. Na execução do ataque, excluímos os conjuntos de quaseidentificadores que estão presentes na base de dados estatística e não estão na amostra. No exemplo condutor, nenhum indivíduo com quaseidentificadores $\{\texttt{ESTADO} = \texttt{SP}, \texttt{SEXO} = \texttt{M}, \texttt{IDADE} = 31 - 40\}$ está na amostra, e portanto essa partição não entra no cálculo da degradação de privacidade esperada pois a degradação de privacidade é 0.

Por fim, calculamos a *degradação de privacidade esperada*.

$$\texttt{degradação-esperada} = \sum_{\substack{i: \\ i \in \{1,2,3\}}} \frac{1}{4} \cdot \frac{n_i/d_i}{4/10}$$

$$\texttt{degradação-esperada} = \frac{1}{4} \cdot \left[\left(\frac{1/1}{4/10}\right) + \left(\frac{2/1}{4/10}\right) + \left(\frac{3/2}{4/10}\right) \right] = 2.8125$$

3.6. Execução dos ataques sobre as bases de dados do Censo Escolar e do PeNSE

Para executar os ataques descritos nas seções anteriores sobre as bases de dados divulgadas para o Censo Escolar da Educação Básica e a Pesquisa Nacional de Saúde do Escolar, utilizamos as divulgações realizadas em 2015[2]. A base de dados do Censo Escolar da Educação Básica possui $48\,536\,347$ registros. Já a base de dados do PeNSE possui $102\,072$ indivíduos.

As colunas utilizadas como quaseidentificadores foram *SEXO*, *ETNIA*, *MÊS DE NASCIMENTO*, *ANO DE NASCIMENTO* e *MUNICIPIO*. Dadas as restrições computacionais para executar os ataques, nos restringimos aos indivíduos presentes na base de dados cujo município é Belo Horizonte, diminuindo a quantidade de registros na base de dados do PeNSE para 2689 e na base de dados do Censo Escolar para $462\,054$.

3.6.1. Ataque de *memberhsip* individual

Para realizar o ataque de *membership* individual, selecionamos todos os indivíduos que são únicos na base de dados do PeNSE, restringindo a seleção a indivíduos cujo município é Belo Horizonte, utilizando os quaseidentificadores citados na seção anterior. O resultado dessa seleção mostra que 125 indivíduos são únicos na amostra, de um total de 2689.

[2]As bases de dados utilizadas estão disponíveis nos seguintes links: PeNSE 2015, Censo Escolar 2015.

ID	ESTADO	SEXO	IDADE	ATRIBUTO SENSÍVEL 1
1	SP	F	31-40	0
2	MG	F	21-30	1
3	MG	F	21-30	1
4	RJ	M	21-30	0
5	RJ	M	21-30	1
6	RJ	M	21-30	0
7	SP	M	31-40	0
8	SP	M	31-40	1
9	SP	M	31-40	0
10	SP	M	31-40	1

Figura 4. Partições geradas na base de dados estatística utilizando as colunas ESTADO, SEXO e IDADE.

.

Em seguida, realizamos seleções na base de dados estatística do Censo Escolar para cada conjunto de valores dos quaseidentificadores dos indivíduos selecionados no passo anterior. Dos 125 indivíduos, 99 foram encontrados na base de dados do Censo escolar [3], e desses 99, 17 deles são únicos na base, i.e., 17 indivíduos foram reidentificados, pois possuem apenas um registro na amostra e na base de dados estatística.

Para esses indivíduos reidentificados, o sucesso a priori é $^{2689}/_{462054}$, o sucesso a posteriori é 1, e a degradação de privacidade é 171.83, ou seja, a divulgação da base de dados do PeNSE aumenta o sucesso do adversário em identificar a presença do indivíduo-alvo na amostra em 171.83 vezes para qualquer um dos 17 indivíduos que puderam ser reidentificados.

3.6.2. Ataque de *membership* coletivo

Para executar o ataque de *membership* coletivo, agrupamos os registros nas bases de dados utilizando as colunas *SEXO*, *ETNIA*, *MÊS DE NASCIMENTO*, *ANO DE NASCIMENTO* e *MUNICIPIO*, gerando um total de 7243 grupos na base de dados do Censo Escolar e 388 grupos na base de dados do PeNSE. Em seguida, calculamos a degradação de privacidade

[3]Relembrando a **Premissa 1** da Seção 3.1, todos os 125 indivíduos deveriam estar presentes na base de dados estatística. Entretanto, os indivíduos que não foram encontrados na base de dados estatística são aqueles que optaram por não responder algumas perguntas do questionário do PeNSE, e portanto não possuem um registro correspondente na base do Censo Escolar quando tratamos apenas dos quaseidentificadores selecionados.

ID	ESTADO	SEXO	IDADE	ATRIBUTO SENSÍVEL 2
A	SP	F	31-40	1
B	MG	F	21-30	0
C	RJ	M	21-30	0
D	RJ	M	21-30	1

Figura 5. Partições geradas na base de dados amostral utilizando as colunas ESTADO, SEXO e IDADE.

.

esperada para cada um dos 388 grupos na base de dados do PeNSE, de acordo com a fórmula descrita na Seção 3.5.

Para esse ataque, a degradação de privacidade esperada é 6.95. Este valor indica que espera-se que a divulgação da base de dados do PeNSE aumente o sucesso do adversário em identificar a presença de um indivíduo-alvo selecionado aleatoriamente na base de dados amostral em 6.95 vezes. Discutiremos na seção seguinte as implicações dos resultados apresentados nesta Seção e na anterior.

3.6.3. Discussão

Como apresentado na Seção 3.6.2, vemos que a técnica de amostragem como método para preservação de privacidade não é eficiente, apresentando riscos à privacidade dos indivíduos presentes na base de dados amostral. Nota-se, entretanto, que o risco é baixo para amostras significativamente menores que a base de dados estatística. Para esses casos, qualquer indivíduo presente na base de dados estatística tem apenas uma pequena probabilidade de ser selecionado para a amostra e, além disso, existe uma probabilidade de ele não ser único na amostra, dificultando o ataque de *membership* e, consequentemente, de reidentificação.

Ademais, vemos pelo ataque de *membership* individual que existem indivíduos na amostra que podem ser unicamente identificados. Para os 17 indivíduos mencionados na Seção 3.6.1, não é suficiente que a amostragem preserve a privacidade de todos os outros indivíduos na base de dados amostral, enquanto que eles estão totalmente vulneráveis ao ataque. Ainda, a Lei Geral de Proteção de Dados prevê a proteção da privacidade para todos os indivíduos presentes na base de dados, não apenas para a grande maioria.

4. Conclusão e trabalhos futuros

Como destacado nesse projeto, a amostragem como técnica de garantia de privacidade, apesar de proteger a grande maioria dos indivíduos presentes na base de dados estatística, não provê garantias de privacidade aos indivíduos presentes na base de dados amostral.

A pesquisa sobre controle de divulgação estatística em bases de dados amostrais

ainda é muito nova, e muitas direções podem ser exploradas para avaliar os riscos aos quais indivíduos presentes nessas bases estão sujeitos.

Uma direção que segue imediatamente dos experimentos realizados na Seção 3.6 é a execução desses ataques para as bases de dados do Censo Escolar e do PeNSE completas, considerando todos os municípios presentes nos dados. Além disso, as análises apresentadas na Seção 3.6 podem ser estendidas para incluir conjuntos diferentes de quaseidentificadores e avaliar quais desses conjuntos apresentam maiores riscos para a privacidade dos indivíduos.

Outra possível direção de pesquisa sobre proteção à privacidade em bases de dados amostrais considera a formalização do modelo de ataque utilizando o arcabouço do campo de Fluxo Quantitativo da Informação. Neste caso, seria necessário criar um modelo teórico para o vazamento de informação causado pela divulgação da base de dados amostral para, assim, derivar resultados teóricos sobre esse tipo de divulgação.

Acreditamos que esse trabalho é um bom ponto de partida para pesquisas futuras no campo de divulgações estatísticas de bases de dados amostrais. Mostramos que é possível realizar ataques sobre bases de dados amostrais e, em alguns casos, a privacidade do indivíduo-alvo é totalmente violada pela divulgação da base amostral.

Referências

[Alvim et al. 2019] Alvim, M., Chatzikokolakis, K., McIver, A., Morgan, C., Palamidessi, C., and Smith, G. (2019). *The Science of Quantitative Information Flow*.

[Dwork 2011] Dwork, C. (2011). A firm foundation for private data analysis. *Commun. ACM*, 54(1):86–95.

[Ghosh et al. 2009] Ghosh, A., Roughgarden, T., and Sundararajan, M. (2009). Universally utility-maximizing privacy mechanisms. In *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing*, STOC '09, page 351–360, New York, NY, USA. Association for Computing Machinery.

[Kifer and Machanavajjhala 2011] Kifer, D. and Machanavajjhala, A. (2011). No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, SIGMOD '11, page 193–204, New York, NY, USA. Association for Computing Machinery.