Técnicas de Privacidade de Dados de Localização

Javam Machado, Eduardo Neto, Manuel Filho

Laboratório de Sistemas e Banco de Dados

05/08/2019

Agenda

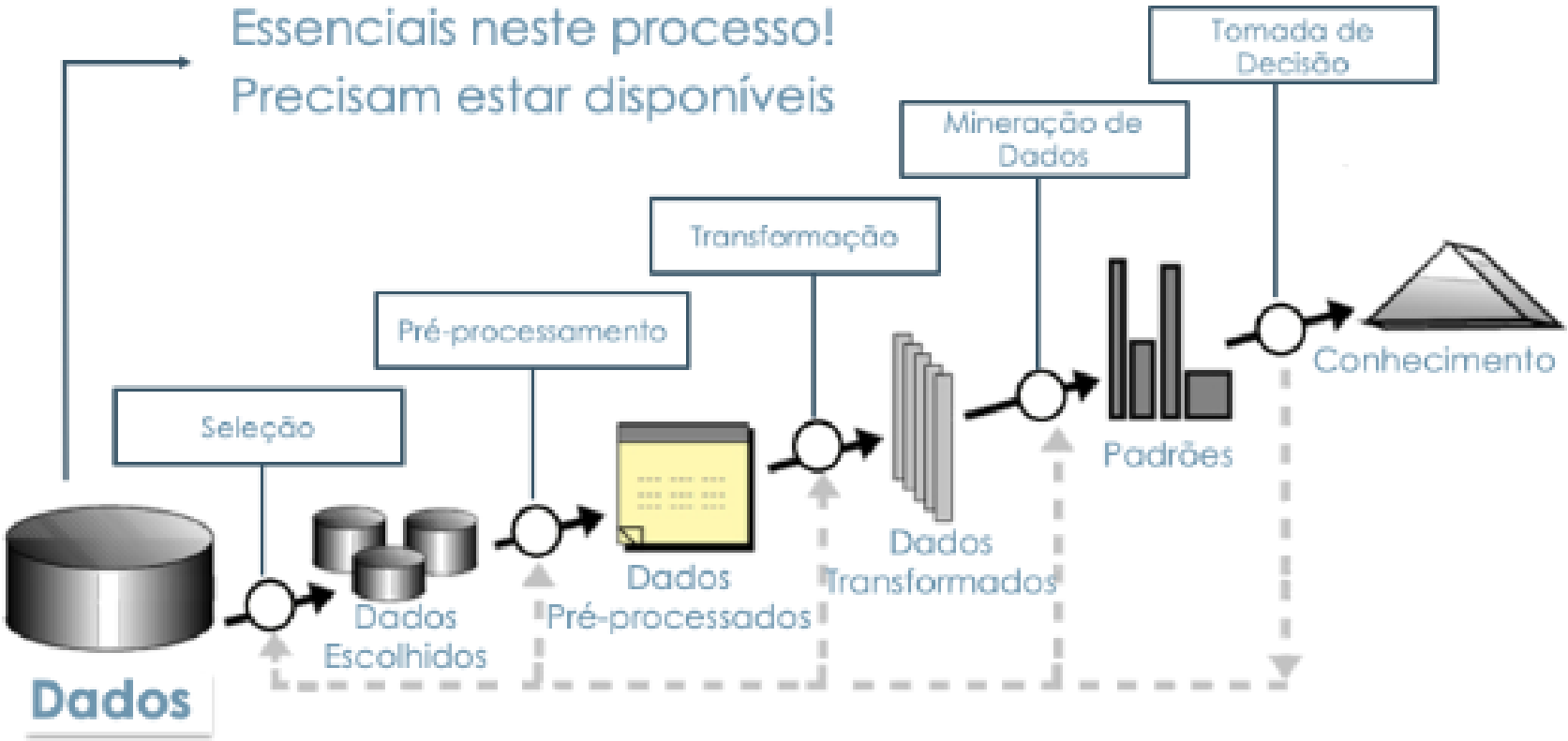
- Privacidade de Dados
  - Motivação
  - Classificação de Atributos
- Maneiras de Proteger Dados Sensíveis
- Técnicas de Anonimização
  - Generalização
  - Supressão
  - Perturbação
- Conhecimento Adversário e Tipos de Ataques
- Modelos de Privacidade Sintáticos
  - k-anonimato
  - l-diversidade
  - δ-presença
- Modelo de Privacidade Diferencial

Privacidade de Dados

Dados, dados, dados...

Source (http://www.agencypja.com/site/assets/files/1826/marketingdata-1.jpg)

Análise e Descoberta de Conhecimento



Benefícios da Disponibilidade de Dados

Disponibilidade de Dados

- Dados privados
  - Não devem ser revelados
  - Derivados de dados pessoais
- Dados que podem ser compartilhados
  - Controle de disponibilidade
  - Necessários a análises, pesquisas, políticas públicas, testes
- Dados abertos
  - Sem restrições de compartilhamento
  - Podem ser utilizados para qualquer objetivo lícito

Privacidade de Dados

E quanto à privacidade dos meus dados?

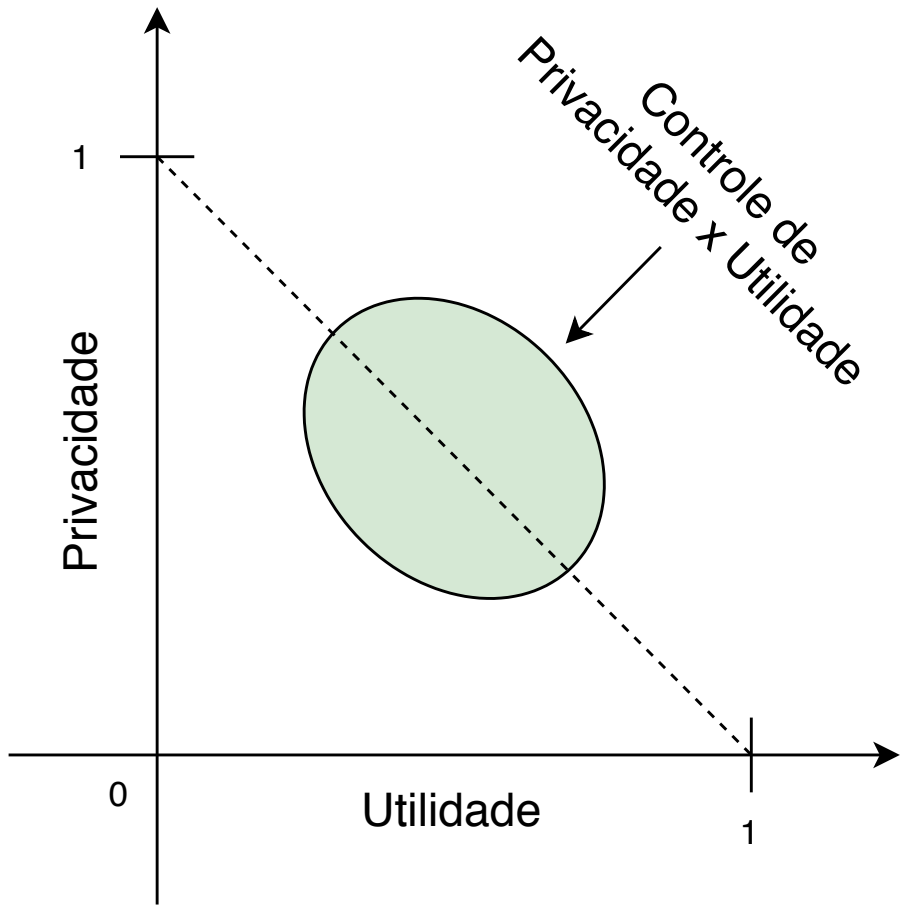
Privacidade

Conceito relacionado a indivíduo

Direito que as pessoas têm em manter um espaço pessoal, sem interferências de outras pessoas ou organizações

Decisão de manter suas informações sob seu exclusivo controle, ou informar, decidindo a quem, quando e onde suas informações estarão disponíveis

Privacidade x Utilidade



Dados Relacionais

atributos

registros

ID	Nome	Idade	Gênero	Endereço	Telefone	Saldo R$
1	Isabela	22	F	Av. I	99998 1324	1.033,25
2	João	25	M	Av. K	99998 1454	814,92
3	Iago	25	M	Av. K	99998 3245	515,09
4	Maria	32	F	Rua J	99998 3465	2.194,79

Classificação de Atributos

- Identificadores explícitos
  - Identificam unicamente indivíduos
  - Sempre removidos antes de qualquer disponibilização

ID	Nome	Idade	Gênero	Endereço	Telefone	Saldo R$
1	Isabela	22	F	Av. I	99998 1324	1.033,25
2	João	25	M	Av. K	99998 1454	814,92
3	Iago	25	M	Av. K	99998 3245	515,09
4	Maria	32	F	Rua J	99998 3465	2.194,79

Classificação de Atributos

- Semi-identificadores
  - Não são identificadores explícitos mas podem identificar indivíduos quando agrupados

ID	Nome	Idade	Gênero	Endereço	Telefone	Saldo R$
1	Isabela	22	F	Av. I	99998 1324	1.033,25
2	João	25	M	Av. K	99998 1454	814,92
3	Iago	25	M	Av. K	99998 3245	515,09
4	Maria	32	F	Rua J	99998 3465	2.194,79

Classificação de Atributos

- Atributos sensíveis
  - Atributos que contêm informações sensíveis sobre os indivíduos

ID	Nome	Idade	Gênero	Endereço	Telefone	Saldo R$
1	Isabela	22	F	Av. I	99998 1324	1.033,25
2	João	25	M	Av. K	99998 1454	814,92
3	Iago	25	M	Av. K	99998 3245	515,09
4	Maria	32	F	Rua J	99998 3465	2.194,79

Maneiras de Proteger Dados Sensíveis

Maneiras de Proteger Dados Sensíveis

- Criptografia
  - Considerada uma das técnicas mais antigas para se proteger dados
  - Utiliza um algoritmo capaz de embaralhar matematicamente dados sensíveis gerando substitutos
  - Substitutos podem ser transformados de volta por meio de chave de acesso
  - **Desvantagem:** utilidade dos dados!

Maneiras de Proteger Dados Sensíveis

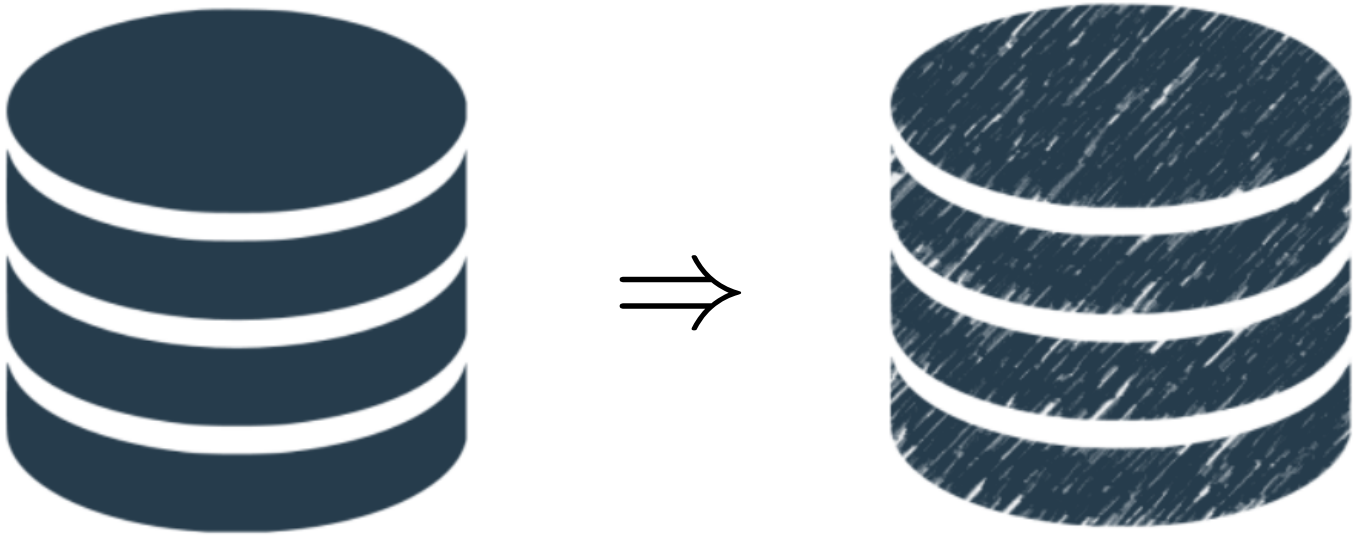
- Tokenização
  - Utilizada quando empresas buscam proteger dados confidenciais já armazenados ou em movimentação para nuvem
  - Gera aleatoriamente um valor de token sem formatação específica
  - Exemplo: Francisco -> F+YCO
  - Ideal para proteger números de cartões de crédito
  - **Desvantagem:** utilidade dos dados!

Maneiras de Proteger Dados Sensíveis

- Anonimização
  - Técnica que modifica dados originais e visa manter a sintaxe e a semântica originais
  - Anonimato: não ser unicamente caracterizado
  - Tem como objetivo o compartilhamento de dados
  - **Desvantagem:** utilidade dos dados!

Técnicas de Anonimização

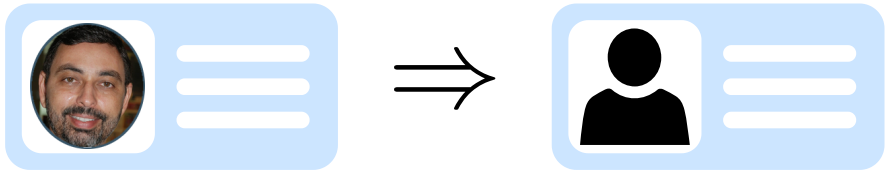
Objetivo da Anonimização



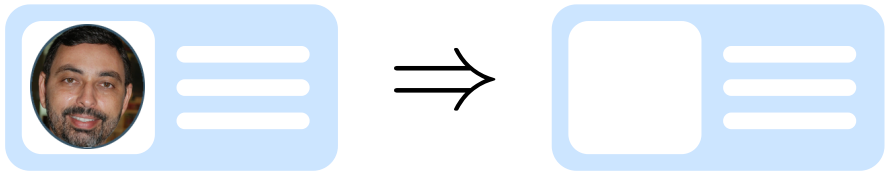
D D'

Técnicas de Anonimização

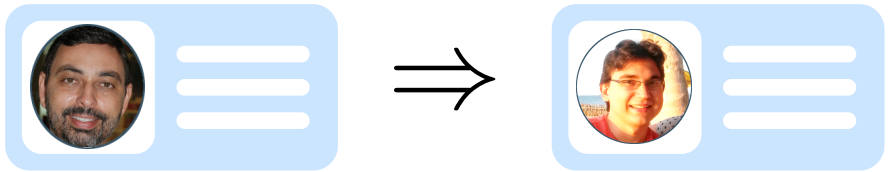
- Generalização



- Supressão



- Perturbação



Generalização

Substituição de valores dos atributos **semi-identificadores** por valores semanticamente semelhantes, porém menos específicos

Idade	⇒	Idade
22		[22-24]
25		[25-29]
25		[25-29]
31		[30-34]

Telefone	⇒	Telefone
99998 1324		99998*
99998 1454		99998*
99998 3245		99998*
99998 3465		99998*

Supressão

Um ou mais valores em um conjunto de dados são removidos ou substituídos por algum valor especial

Idade
22
25
25
31

$\Rightarrow$

Idade

Telefone
99998 1324
99998 1454
99998 3245
99998 3465

$\Rightarrow$

Telefone
*
*
*
*

Supressão

- **Supressão de Registro:** Um registro é removido inteiramente do conjunto de dados
- **Supressão de Valor:** Remoção de todas as instâncias de um determinado valor (ou intervalo) de um atributo
- **Supressão de Células:** Apenas algumas instâncias de valores de um atributo são removidas

Perturbação

- Substituição de valores de atributos semi-identificadores originais por valores fictícios
- Informações estatísticas calculadas a partir dos dados originais não diferenciam significativamente de informações estatísticas calculadas a partir dos dados perturbados
- Não preserva a veracidade dos dados

Perturbação

- **Adição de Ruído:** Consiste em substituir um valor original de atributo "v" por "v+r", onde "r" é um valor, denominado **ruído**
- **Permutação de Dados:** Consiste em permutar dois valores do mesmo atributo de registros diferentes
- **Geração de Dados Sintéticos:** Consiste em duas etapas:
  1. Gerar um modelo estatístico a partir do conjunto de dados
  2. Gerar dados sintéticos a partir do modelo estatístico

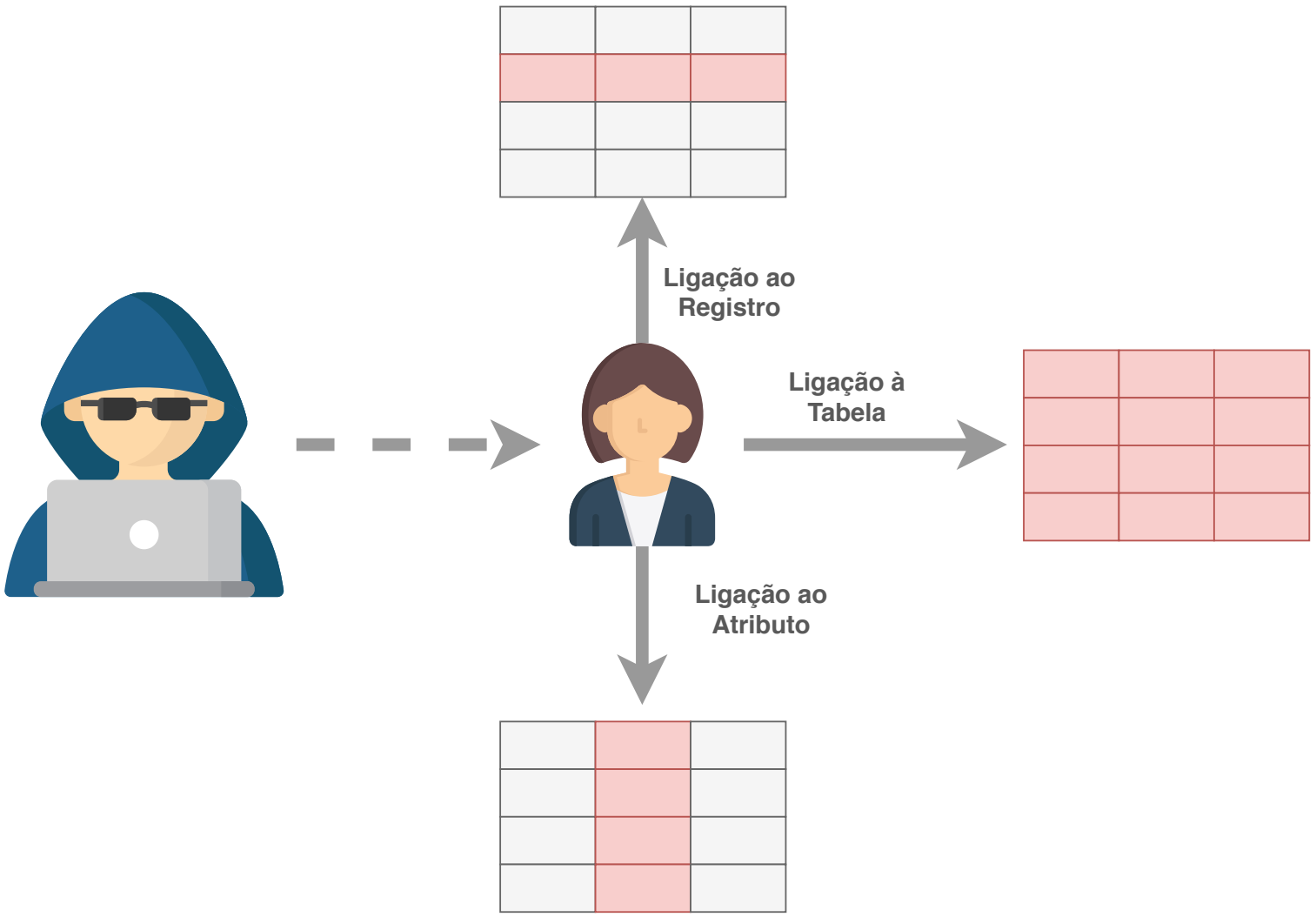
Conhecimento Adversário e Tipos de Ataques

Conhecimento adversário

- Conhecimento previamente adquirido de fontes externas

Esse conhecimento é imprevisível

Tipos de Ataques à Privacidade

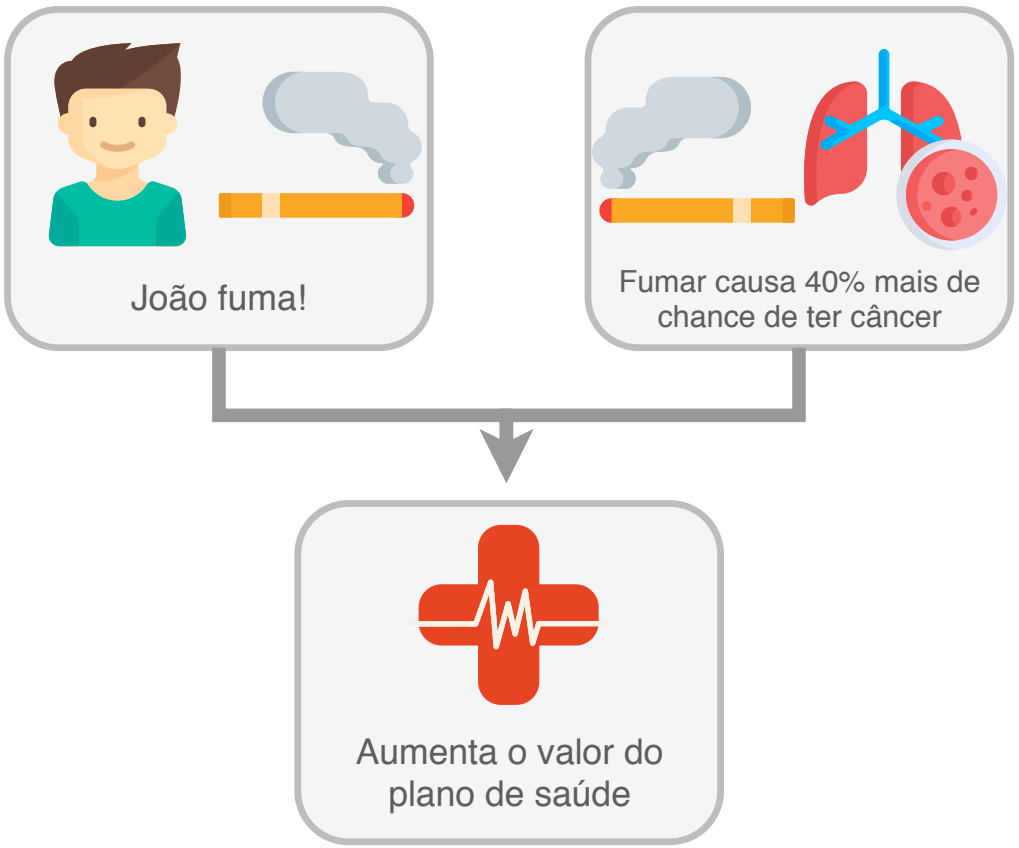


Tipos de Ataques à Privacidade

- Ataque de Ligação ao Registro
  - Objetivo: Re-identificar o registro de um indivíduo
- Ataque de Ligação ao Atributo
  - Objetivo: Inferir atributos sensíveis mesmo sem re-identificação
- Ataque de Ligação à Tabela
  - Objetivo: Inferir a presença ou a ausência de um indivíduo no conjunto de dados

Tipos de Ataques à Privacidade

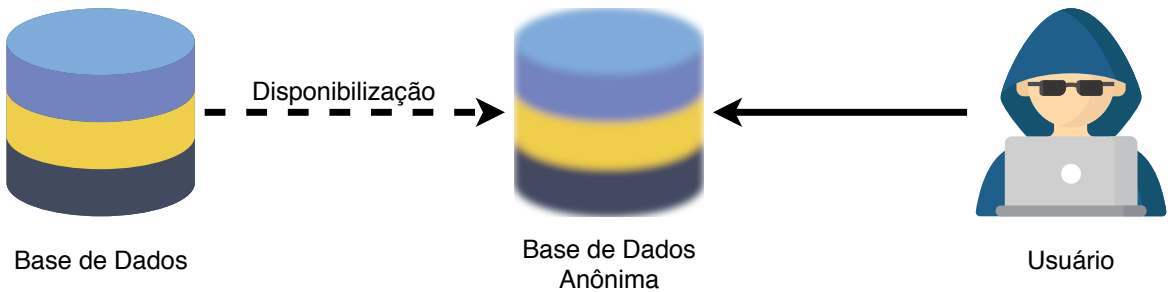
- Ataque Probabilístico
  - Objetivo: Mudar a forma de pensar (ou não) acerca de um indivíduo após ter acessado informações externas



Modelos de Privacidade Sintáticos

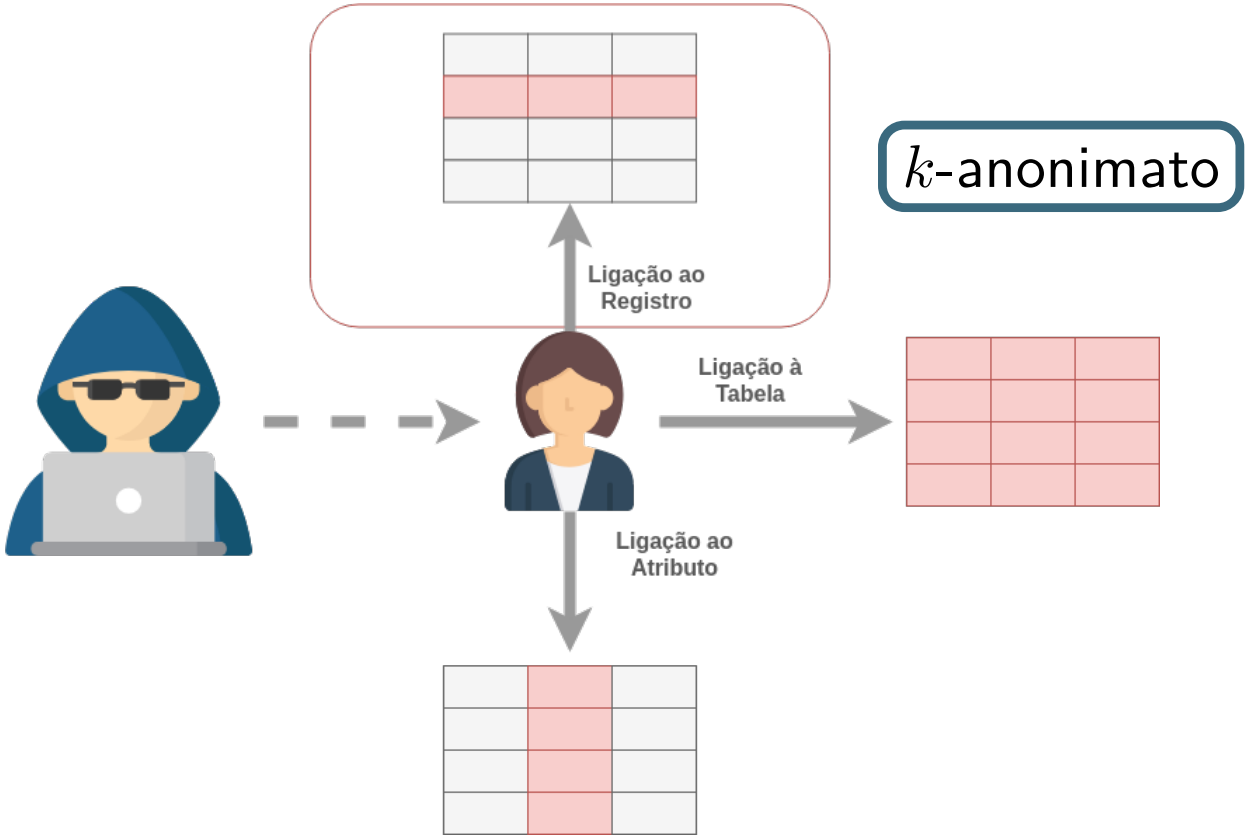
Modelos de Privacidade Sintáticos

- Dados são disponibilizados
- Generalização e/ou supressão (de modo geral)
- Atendem a uma condição sintática



k-anonimato

- Ataques de ligação ao registro



k-anonimato

O registro de um indivíduo não deve ser identificado com probabilidade maior que $\frac{1}{k}$

k-anonimato

O registro de um indivíduo não deve ser identificado com probabilidade maior que $\frac{1}{k}$

$\Updownarrow$

Tabela é logicamente particionada em classes de equivalência

k-anonimato

O registro de um indivíduo não deve ser identificado com probabilidade maior que $\frac{1}{k}$

$\Updownarrow$

Tabela é logicamente particionada em classes de equivalência

$\Updownarrow$

As classes de equivalência devem ter tamanho mínimo de k

k-anonimato

O registro de um indivíduo não deve ser identificado com probabilidade maior que $\frac{1}{k}$

$\Updownarrow$

Tabela é logicamente particionada em classes de equivalência

$\Updownarrow$

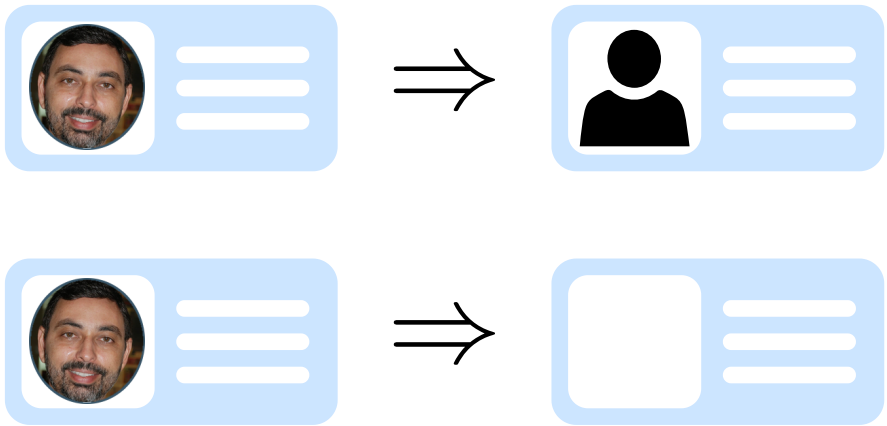
As classes de equivalência devem ter tamanho mínimo de k

$\Updownarrow$

Para cada registro, devem existir, pelo menos, k-1 registros iguais a ele

Como anonimizar?

- Generalização e/ou Supressão



Exemplo

- 2-anonimato com Supressão

Cidade	Idade	Renda
Fortaleza	20	1800
Sobral	25	920
Quixadá	20	1800
Sobral	22	2200

Exemplo

- 2-anonimato com Supressão

Cidade	Idade	Renda
Fortaleza	20	1800
Sobral	25	920
Quixadá	20	1800
Sobral	22	2200

$\Rightarrow$

Cidade	Idade	Renda
*	20	1800
Sobral	*	*
*	20	1800
Sobral	*	*

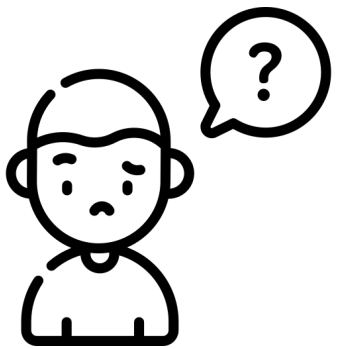
Exemplo

- 2-anonimato com Supressão

Cidade	Idade	Renda
Fortaleza	20	1800
Sobral	25	920
Quixadá	20	1800
Sobral	22	2200

$\Rightarrow$

Cidade	Idade	Renda
*	20	1800
Sobral	*	*
*	20	1800
Sobral	*	*



Esse dado parece bom para análise?

Exemplo

- 2-anonimato com Generalização

Profissão	Idade	Cidade	Doença
Engenheiro	29	Fortaleza	Chinkungunya
Médico	29	Fortaleza	Chinkungunya
Médico	29	Fortaleza	Hepatite C
Programador	32	Fortaleza	AIDS
Programador	32	Quixadá	AIDS
Dentista	35	Quixadá	AIDS

Exemplo

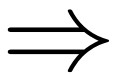
- 2-anonimato com Generalização

Profissão	Idade	Cidade	Doença
Engenheiro	29	Fortaleza	Chinkungunya
Médico	29	Fortaleza	Chinkungunya
Médico	29	Fortaleza	Hepatite C
Programador	32	Fortaleza	AIDS
Programador	32	Quixadá	AIDS
Dentista	35	Quixadá	AIDS

Exemplo

- 2-anonimato com Generalização

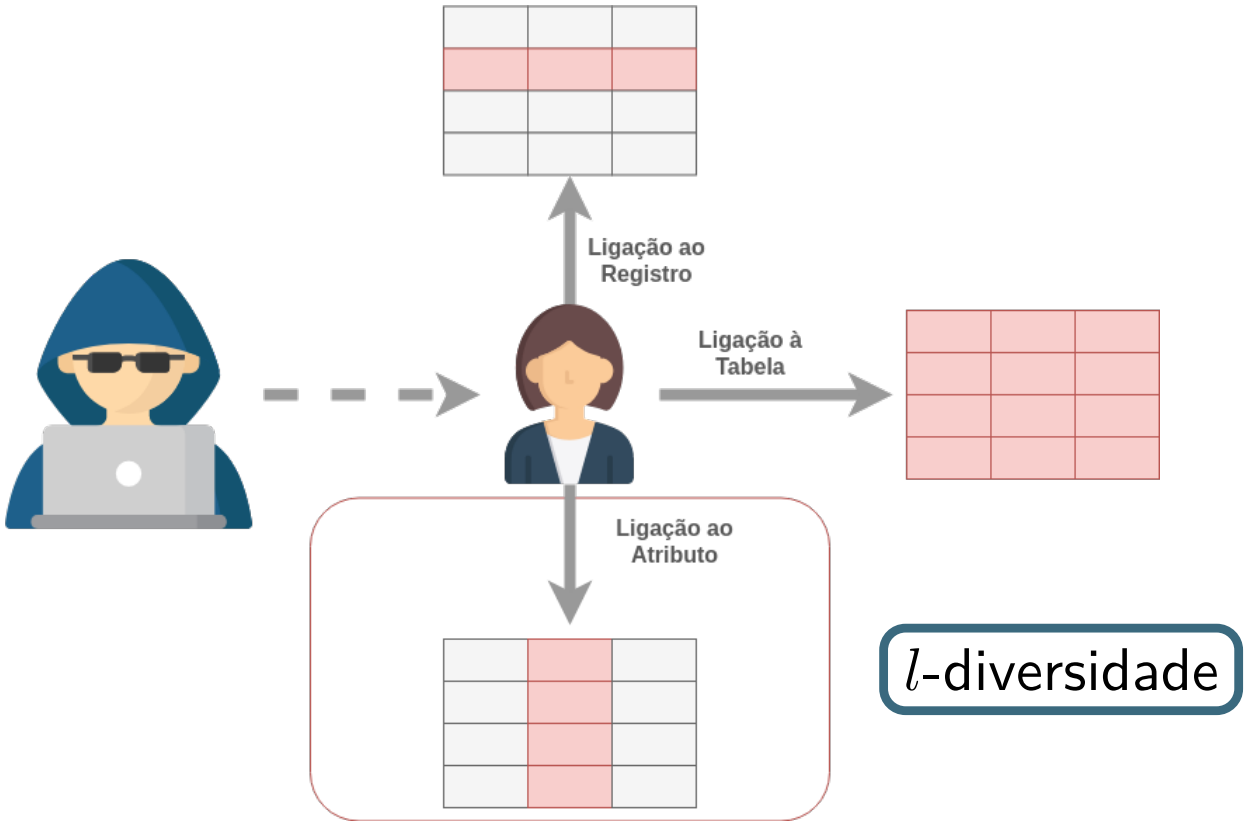
Profissão	Idade	Cidade	Doença
Engenheiro	29	Fortaleza	Chinkungunya
Médico	29	Fortaleza	Chinkungunya
Médico	29	Fortaleza	Hepatite C
Programador	32	Fortaleza	AIDS
Programador	32	Quixadá	AIDS
Dentista	35	Quixadá	AIDS



Profissão	Idade	Cidade	Doença
Profissional	[25-35]	CE	Chinkungunya
Médico	29	Fortaleza	Chinkungunya
Médico	29	Fortaleza	Hepatite C
Programador	32	CE	AIDS
Programador	32	CE	AIDS
Profissional	[25-35]	CE	AIDS

l-diversidade

- Ataques de ligação ao atributo



l-diversidade

- Atua como complemento do k-anonimato
- Protege contra ataques de ligação ao **atributo**
- k-anonimato + l-diversidade
  - Proteção contra ataques de ligação ao **registro** e ligação ao **atributo**

l-diversidade

Cada classe de equivalência deve possuir pelo menos l valores distintos para o atributo sensível.

Exemplo

- 2-diversidade

Profissão	Idade	Cidade	Doença
Profissional	[25-35]	CE	Chikungunya
Médico	29	Fortaleza	Chikungunya
Médico	29	Fortaleza	Hepatite C
Programador	32	CE	AIDS
Programador	32	CE	AIDS
Profissional	[25-35]	CE	AIDS

✓ 2-anônimo

Exemplo

- 2-diversidade

Profissão	Idade	Cidade	Doença
Profissional	[25-35]	CE	Chikungunya
Médico	29	Fortaleza	Chikungunya
Médico	29	Fortaleza	Hepatite C
Programador	32	CE	AIDS
Programador	32	CE	AIDS
Profissional	[25-35]	CE	AIDS

✓ 2-diverso

Exemplo

- 2-diversidade

Profissão	Idade	Cidade	Doença
Profissional	[25-35]	CE	Chikungunya
Médico	29	Fortaleza	Chikungunya
Médico	29	Fortaleza	Hepatite C
Programador	32	CE	AIDS
Programador	32	CE	AIDS
Profissional	[25-35]	CE	AIDS

✓ 2-diverso

Exemplo

- 2-diversidade

Profissão	Idade	Cidade	Doença
Profissional	[25-35]	CE	Chikungunya
Médico	29	Fortaleza	Chikungunya
Médico	29	Fortaleza	Hepatite C
Programador	32	CE	AIDS
Programador	32	CE	AIDS
Profissional	[25-35]	CE	AIDS

⚠ Problema

Exemplo

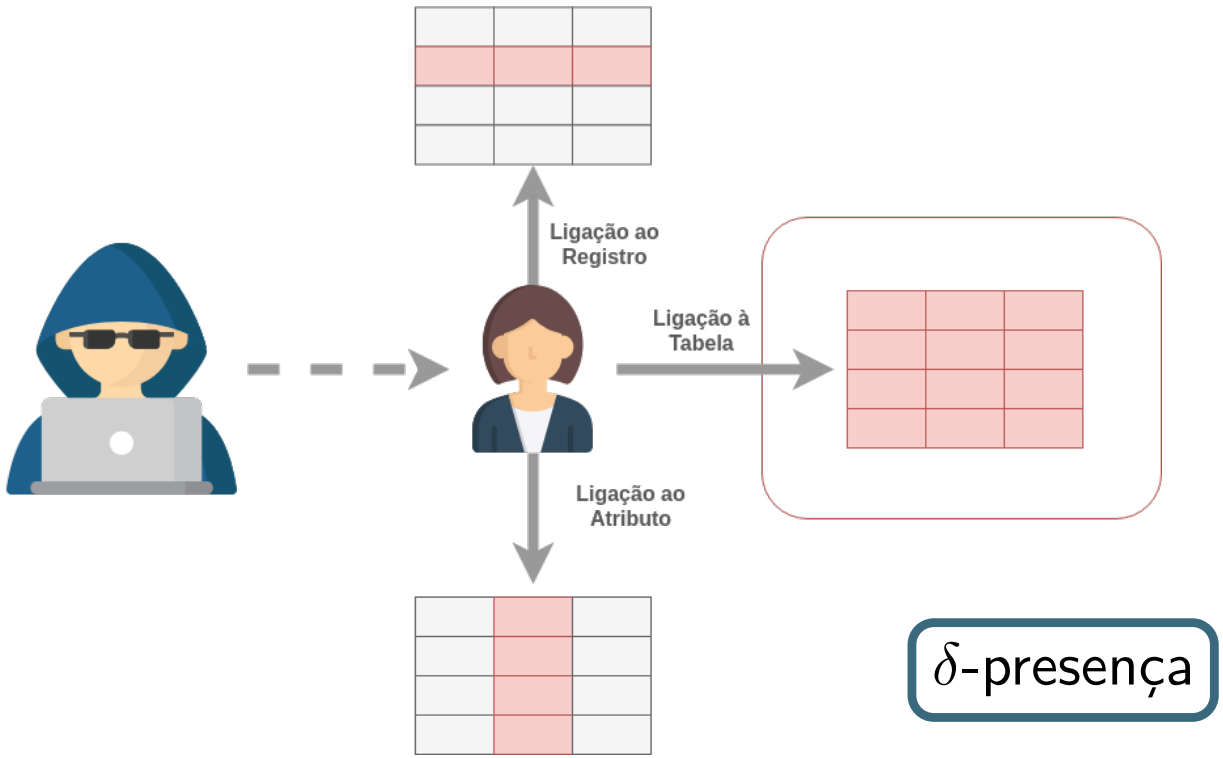
- 2-diversidade

Profissão	Idade	Cidade	Doença
Profissional	[25-35]	CE	Chikungunya
Médico	29	Fortaleza	Chikungunya
Médico	29	Fortaleza	Hepatite C
Programador	32	CE	AIDS
Programador	32	CE	***Gripe***
Profissional	[25-35]	CE	AIDS

✓ 2-diverso

δ-presença

- Ataques de ligação à tabela



δ-presença

- Define $\delta = (\delta_{min}, \delta_{max})$
  - Limites de probabilidade de inferir a presença de um indivíduo na tabela

Exemplo

Subconjunto de Pesquisa (T)

ID	CEP	Idade	País
	47903	59	Canadá
	47906	42	EUA
	47633	63	Peru
	48972	47	Bulgária
	48970	52	França

Exemplo

Subconjunto de Pesquisa (T)

ID	CEP	Idade	País
	47903	59	Canadá
	47906	42	EUA
	47633	63	Peru
	48972	47	Bulgária
	48970	52	França

Exemplo

Subconjunto de Pesquisa (T)

ID	CEP	Idade	País
	47*	59	Canadá
	47*	42	EUA
	47*	63	Peru
	48972	47	Bulgária
	48970	52	França

Exemplo

Subconjunto de Pesquisa (T)

ID	CEP	Idade	País
	47*	*	Canadá
	47*	*	EUA
	47*	*	Peru
	48972	47	Bulgária
	48970	52	França

Exemplo

Subconjunto de Pesquisa (T)

ID	CEP	Idade	País
	47*	*	América
	47*	*	América
	47*	*	América
	48972	47	Bulgária
	48970	52	França

Exemplo

Subconjunto de Pesquisa (T)

ID	CEP	Idade	País
	47*	*	América
	47*	*	América
	47*	*	América
	48972	47	Bulgária
	48970	52	França

Exemplo

Subconjunto de Pesquisa (T)

ID	CEP	Idade	País
	47*	*	América
	47*	*	América
	47*	*	América
	48*	47	Bulgária
	48*	52	França

Exemplo

Subconjunto de Pesquisa (T)

ID	CEP	Idade	País
	47*	*	América
	47*	*	América
	47*	*	América
	48*	*	Bulgária
	48*	*	França

Exemplo

Subconjunto de Pesquisa (T)

ID	CEP	Idade	País
	47*	*	América
	47*	*	América
	47*	*	América
	48*	*	Europa
	48*	*	Europa

$\subseteq$

Dado Público (E)

ID	Nome	CEP	Idade	País
a	Alice	47906	35	EUA
b	Bob	47903	59	Canadá
c	Christine	47906	42	EUA
d	Dirk	47630	18	Brasil
e	Eunice	47630	22	Brasil
f	Frank	47633	63	Peru
g	Gail	48973	33	Espanha
h	Harry	48972	47	Bulgária
i	Iris	48970	52	França

Exemplo

Subconjunto de Pesquisa (T)

ID	CEP	Idade	País
b	47*	*	América
c	47*	*	América
f	47*	*	América
h	48*	*	Europa
i	48*	*	Europa

$\subseteq$

Dado Público (E)

ID	Nome	CEP	Idade	País
a	Alice	47906	35	EUA
b	Bob	47903	59	Canadá
c	Christine	47906	42	EUA
d	Dirk	47630	18	Brasil
e	Eunice	47630	22	Brasil
f	Frank	47633	63	Peru
g	Gail	48973	33	Espanha
h	Harry	48972	47	Bulgária
i	Iris	48970	52	França

Exemplo

Subconjunto de Pesquisa (T)

ID	CEP	Idade	País
b	47*	*	América
c	47*	*	América
f	47*	*	América
h	48*	*	Europa
i	48*	*	Europa

$\subseteq$

Dado Público (E)

ID	Nome	CEP	Idade	País
a	Alice	47906	35	EUA
b	Bob	47903	59	Canadá
c	Christine	47906	42	EUA
d	Dirk	47630	18	Brasil
e	Eunice	47630	22	Brasil
f	Frank	47633	63	Peru
g	Gail	48973	33	Espanha
h	Harry	48972	47	Bulgária
i	Iris	48970	52	França

Exemplo

Subconjunto de Pesquisa (T)

ID	CEP	Idade	País
b	47*	*	América
c	47*	*	América
f	47*	*	América
h	48*	*	Europa
i	48*	*	Europa

$\subseteq$

Dado Público (E)

ID	Nome	CEP	Idade	País
a	Alice	47906	35	EUA
b	Bob	47903	59	Canadá
c	Christine	47906	42	EUA
d	Dirk	47630	18	Brasil
e	Eunice	47630	22	Brasil
f	Frank	47633	63	Peru
g	Gail	48973	33	Espanha
h	Harry	48972	47	Bulgária
i	Iris	48970	52	França

Exemplo

Subconjunto de Pesquisa (T)

ID	CEP	Idade	País
b	47*	*	América
c	47*	*	América
f	47*	*	América
h	48*	*	Europa
i	48*	*	Europa

$\subseteq$

Dado Público (E)

ID	Nome	CEP	Idade	País
a	Alice	47906	35	EUA
b	Bob	47903	59	Canadá
c	Christine	47906	42	EUA
d	Dirk	47630	18	Brasil
e	Eunice	47630	22	Brasil
f	Frank	47633	63	Peru
g	Gail	48973	33	Espanha
h	Harry	48972	47	Bulgária
i	Iris	48970	52	França

Qual a probabilidade de **Alice** estar no dado anonimizado?

Exemplo

Subconjunto de Pesquisa (T)

ID	CEP	Idade	País
b	47*	*	América
c	47*	*	América
f	47*	*	América
h	48*	*	Europa
i	48*	*	Europa

$\subseteq$

Dado Público (E)

ID	Nome	CEP	Idade	País
a	Alice	47906	35	EUA
b	Bob	47903	59	Canadá
c	Christine	47906	42	EUA
d	Dirk	47630	18	Brasil
e	Eunice	47630	22	Brasil
f	Frank	47633	63	Peru
g	Gail	48973	33	Espanha
h	Harry	48972	47	Bulgária
i	Iris	48970	52	França

$$Prob(Alice \in T|E) = ?$$

Exemplo

Subconjunto de Pesquisa (T)

ID	CEP	Idade	País
b	47*	*	América
c	47*	*	América
f	47*	*	América
h	48*	*	Europa
i	48*	*	Europa

$\subseteq$

Dado Público (E)

ID	Nome	CEP	Idade	País
a	Alice	47906	35	EUA
b	Bob	47903	59	Canadá
c	Christine	47906	42	EUA
d	Dirk	47630	18	Brasil
e	Eunice	47630	22	Brasil
f	Frank	47633	63	Peru
g	Gail	48973	33	Espanha
h	Harry	48972	47	Bulgária
i	Iris	48970	52	França

$$Prob(Alice \in T|E) = \frac{3}{6} = 0.50 = 50\%$$

Exemplo

Subconjunto de Pesquisa (T)

ID	CEP	Idade	País
b	47*	*	América
c	47*	*	América
f	47*	*	América
h	48*	*	Europa
i	48*	*	Europa

$\subseteq$

Dado Público (E)

ID	Nome	CEP	Idade	País
a	Alice	47906	35	EUA
b	Bob	47903	59	Canadá
c	Christine	47906	42	EUA
d	Dirk	47630	18	Brasil
e	Eunice	47630	22	Brasil
f	Frank	47633	63	Peru
g	Gail	48973	33	Espanha
h	Harry	48972	47	Bulgária
i	Iris	48970	52	França

Qual a probabilidade de **Harry** estar no dado anonimizado?

Exemplo

Subconjunto de Pesquisa (T)

ID	CEP	Idade	País
b	47*	*	América
c	47*	*	América
f	47*	*	América
h	48*	*	Europa
i	48*	*	Europa

$\subseteq$

Dado Público (E)

ID	Nome	CEP	Idade	País
a	Alice	47906	35	EUA
b	Bob	47903	59	Canadá
c	Christine	47906	42	EUA
d	Dirk	47630	18	Brasil
e	Eunice	47630	22	Brasil
f	Frank	47633	63	Peru
g	Gail	48973	33	Espanha
h	Harry	48972	47	Bulgária
i	Iris	48970	52	França

$$Prob(Harry \in T|E) =?$$

Exemplo

Subconjunto de Pesquisa (T)

ID	CEP	Idade	País
b	47*	*	América
c	47*	*	América
f	47*	*	América
h	48*	*	Europa
i	48*	*	Europa

$\subseteq$

Dado Público (E)

ID	Nome	CEP	Idade	País
a	Alice	47906	35	EUA
b	Bob	47903	59	Canadá
c	Christine	47906	42	EUA
d	Dirk	47630	18	Brasil
e	Eunice	47630	22	Brasil
f	Frank	47633	63	Peru
g	Gail	48973	33	Espanha
h	Harry	48972	47	Bulgária
i	Iris	48970	52	França

$$Prob(Harry \in T|E) = \frac{2}{3} \approx 0.66 \approx 66\%$$

Exemplo

Subconjunto de Pesquisa (T)

ID	CEP	Idade	País
b	47*	*	América
c	47*	*	América
f	47*	*	América
h	48*	*	Europa
i	48*	*	Europa

$\subseteq$

Dado Público (E)

ID	Nome	CEP	Idade	País
a	Alice	47906	35	EUA
b	Bob	47903	59	Canadá
c	Christine	47906	42	EUA
d	Dirk	47630	18	Brasil
e	Eunice	47630	22	Brasil
f	Frank	47633	63	Peru
g	Gail	48973	33	Espanha
h	Harry	48972	47	Bulgária
i	Iris	48970	52	França

$$\left.\begin{array}{l}\delta_{min} = 0.50 \\ \delta_{max} = 0.66\end{array}\right\} = (0.50, 0.66)\text{-Presença}$$

Modelo de Privacidade Diferencial

Modelo de Privacidade Diferencial

- É um **modelo matemático** e não uma condição sintática.

$$log\left(\frac{Pr(M(D_1) = O)}{Pr(M(D_2) = O)}\right) \leq \varepsilon$$

- A diferença entre as probabilidades de uma consulta retornar o mesmo resultado em dois conjuntos de dados é limitada pelo parâmetro ε (*budget*).

Conjunto de dados vizinhos

- Pares de entradas que diferem em apenas uma tupla
  - Ausência ou presença
  - Valor

ID	Peso (Kg)	Altura (m)
1	87,2	1,7
2	81,2	1,62
3	74,2	1,75
4	60	1,61
5	78,5	1,58

Tabela: D_1

ID	Peso (Kg)	Altura (m)
1	87,2	1,7
2	81,2	1,62
4	60	1,61
5	78,5	1,58

Tabela: D_2

Privacidade Diferencial

- Disponibiliza, de maneira geral, informações estatísticas sobre conjuntos de dados



- Hospital
- Pesquisas médicas

- Censo
- Economistas

- Google
- Sistemas de recomendação

Privacidade Diferencial

- Disponibiliza, de maneira geral, informações estatísticas sobre conjuntos de dados



- Hospital
- Pesquisas médicas

- Censo
- Economistas

- Google
- Sistemas de recomendação

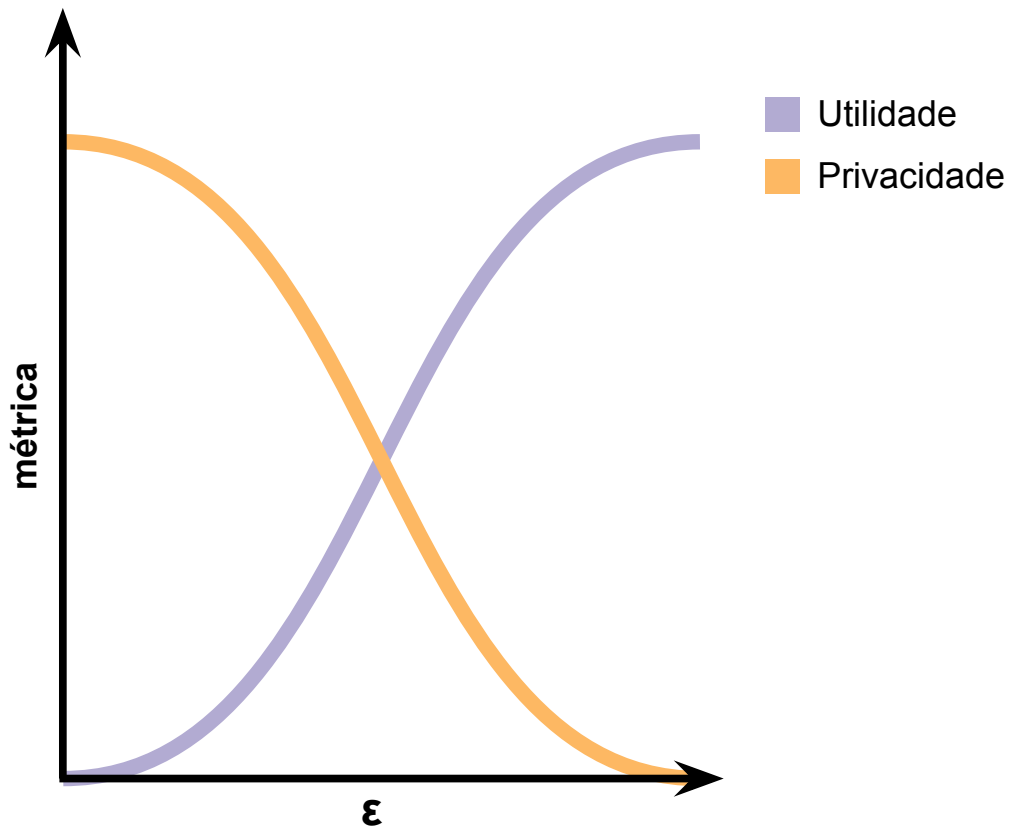
Informações estatísticas de consultas de agregação são muito usadas em processos de mineração.

Limite de privacidade ε

- Parâmetro: ε-privacidade diferencial
- $log\left(\frac{Pr(M(D)=O)}{Pr(M(D')=O)}\right) \leq \varepsilon$
- Controla o grau de indistinguibilidade
- Quanto menor o ε, menor deve ser a diferença de probabilidades
- Dependente da consulta
- Valores recomendados: 0.01, 0.1, $ln2$ e $ln3$

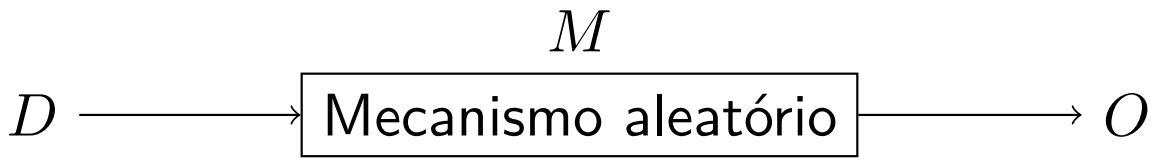
Limite de privacidade ε – *Trade-off*

- $log\left(\frac{Pr(M(D)=O)}{Pr(M(D')=O)}\right) \leq \varepsilon$
- $\Downarrow$ budget ε $\Longrightarrow$ $\Uparrow$ privacidade



Mecanismo

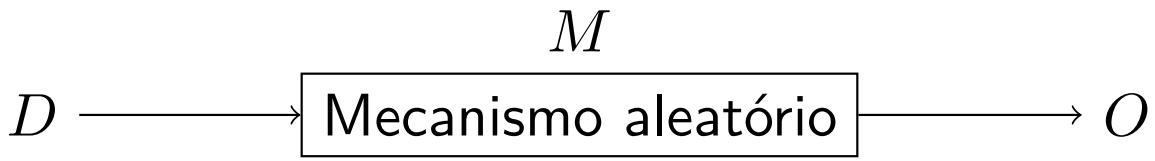
Ideia geral:



- Qualquer saída O de M é produzida com **quase** a mesma probabilidade, não importando se um indivíduo específico está na base de dados D.

Mecanismo

Ideia geral:

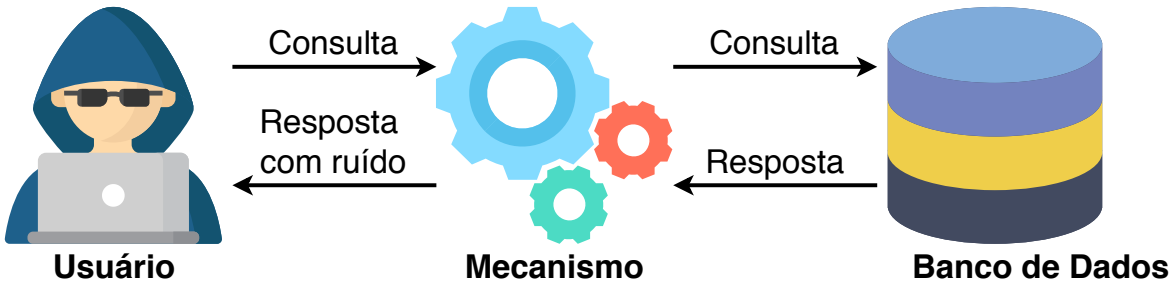


- Qualquer saída O de M é produzida com **quase** a mesma probabilidade, não importando se um indivíduo específico está na base de dados D.
- Um mecanismo M satisfaz ε-privacidade diferencial sse:

$$log\left(\frac{Pr(M(D) = O)}{Pr(M(D') = O)}\right) \leq \varepsilon$$

- Para quaisquer dois datasets D e D' vizinhos e todas as possíveis saídas O

Mecanismo – *Overview* no fluxo da PD



- Consultas: count, sum, avg, min, max

Mecanismo

A definição:

$$log\left(\frac{Pr(M(D)=O)}{Pr(M(D')=O)}\right) \leq \varepsilon$$

Também é comumente apresentada da seguinte forma:

$$Pr(M(D)=O) \leq exp(\varepsilon)Pr(M(D')=O)$$

A diferença entre as probabilidades de uma consulta retornar o mesmo resultado em dois conjuntos de dados é limitada pelo parâmetro ε.

Bibliografia I

- Aggarwal, Charu C e S Yu Philip (2008). "A framework for condensation-based anonymization of string data". Em: *Data Mining and Knowledge Discovery* 16.3, pp. 251–275.
- Bayardo, Roberto J e Rakesh Agrawal (2005). "Data privacy through optimal k-anonymization". Em: *21st International conference on data engineering (ICDE'05)*, pp. 217–228.
- Brito, Felipe e Javam Machado (2017). "Preservação de Privacidade de Dados: Fundamentos, Técnicas e Aplicações". Em: p. 40. ISBN: 978-85-7669-374-1.
- Domingo-Ferrer, Josep e Vicenc Torra (2001). "A quantitative comparison of disclosure control methods for microdata". Em: *Confidentiality, disclosure and data access: theory and practical applications for statistical agencies*, pp. 111–134.
- Dwork, C (2008). *Differential privacy: a survey of results. InInternational conference on theory and applications of models of computation (pp. 1–19)*.

Bibliografia II


Dwork, Cynthia (2006). "Differential Privacy". Em: *33rd International Colloquium on Automata, Languages and Programming*. Venice, Italy, pp. 1–12.


Fung, Benjamin C.M., Ke Wang, Ada Wai-Chee Fu e Philip S. Yu (2010). *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques*. Ed. por Vipin Kumar. 1st. ISBN 978-1-4200-9148-9. Chapman & Hall/CRC. ISBN: 1420091484, 9781420091489.


Machanavajjhala, Ashwin, Johannes Gehrke, Daniel Kifer e Muthuramakrishnan Venkitasubramaniam (2006). "l-diversity: Privacy beyond k-anonymity". Em: *22nd International Conference on Data Engineering (ICDE'06)*. IEEE, pp. 24–24.


Meyerson, Adam e Ryan Williams (2004). "On the complexity of optimal k-anonymity". Em: *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, pp. 223–228.

Bibliografia III


Nergiz, Mehmet Ercan, Maurizio Atzori e Chris Clifton (2007). "Hiding the Presence of Individuals from Shared Databases". Em: *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*. SIGMOD '07. Beijing, China: ACM, pp. 665–676. ISBN: 978-1-59593-686-8. DOI: 10.1145/1247480.1247554. URL: http://doi.acm.org/10.1145/1247480.1247554.


Sweeney, Latanya (2002). "k-anonymity: A model for protecting privacy". Em: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05, pp. 557–570.


Tan, Vincent Yan Fu e See-Kiong Ng (2007). "Generic probability density function reconstruction for randomization in privacy-preserving data mining". Em: *International Workshop on Machine Learning and Data Mining in Pattern Recognition*. Springer, pp. 76–90.


Wang, Ke, Benjamin CM Fung e Philip S Yu (2005). "Template-based privacy preservation in classification problems". Em: *Fifth IEEE International Conference on Data Mining (ICDM'05)*. IEEE, 8–pp.


Wong, Raymond Chi-Wing e Ada Wai-Chee Fu (2010). "Privacy-preserving data publishing: An overview". Em: *Synthesis Lectures on Data Management* 2.1, pp. 1–138.