Estudo de Casos sobre Privacidade e Transparência na Publicação de Dados

Gabriel Henrique Nunes

EVCOMP 2020

Departamento de Ciência da Computação
Universidade Federal de Minas Gerais

17 de Fevereiro de 2020

INSCRYPT
Laboratory of Information Security, Cryptography, Privacy, and Transparency

Introdução

Transparência

A Importância da Publicação de Dados

The actual Enumeration shall be made within three Years after the first Meeting of the Congress of the United States, and within every subsequent Term of ten Years, in such Manner as they shall by Law direct.

A Importância da Publicação de Dados

Planejamento

- Políticas públicas

A Importância da Publicação de Dados

Planejamento

- Políticas públicas
  - Representação política

A Importância da Publicação de Dados

Planejamento

- Políticas públicas
  - Representação política
  - Distribuição de fundos

A Importância da Publicação de Dados

Planejamento

- Políticas públicas
  - Representação política
  - Distribuição de fundos
- Investimentos privados

A Importância da Publicação de Dados

Planejamento

- Políticas públicas
  - Representação política
  - Distribuição de fundos
- Investimentos privados

Pesquisas científicas

- Demografia
- Economia

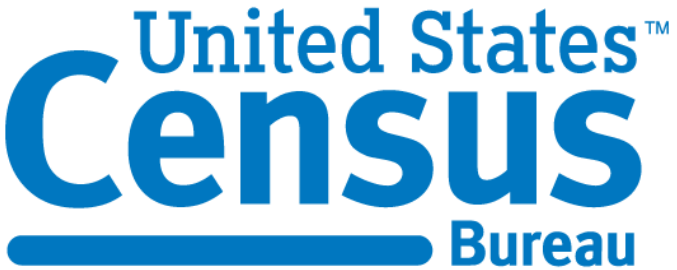
A Importância da Publicação de Dados

Planejamento

- Políticas públicas
  - Representação política
  - Distribuição de fundos
- Investimentos privados

Pesquisas científicas

- Demografia
- Economia



US Code Titles 13 & 14

A Importância da Publicação de Dados

Planejamento

- Políticas públicas
  - Representação política
  - Distribuição de fundos
- Investimentos privados

Pesquisas científicas

- Demografia
- Economia

Legislação específica
(1968)

Lei de Acesso à Informação
(2011)

A Importância da Publicação de Dados

Planejamento

- Políticas públicas
  - Representação política
  - Distribuição de fundos
- Investimentos privados

Pesquisas científicas

- Demografia
- Economia

Princípios Fundamentais das Estatísticas Oficiais

(2014)

Introdução

Privacidade

O Problema da Publicação de Dados

Princípio 6. Os dados individuais coletados pelos órgãos estatísticos para a elaboração de estatísticas, sejam eles referentes a pessoas físicas ou jurídicas, devem ser estritamente confidenciais e utilizados exclusivamente para fins estatísticos.

UNITED NATIONS

O Problema da Publicação de Dados

Princípio 6. Os dados individuais coletados pelos órgãos estatísticos para a elaboração de estatísticas, sejam eles referentes a pessoas físicas ou jurídicas, devem ser estritamente confidenciais e utilizados exclusivamente para fins estatísticos.

IBGE
Instituto Brasileiro de Geografia e Estatística

United States™ Census Bureau

O Problema da Publicação de Dados

A Importância da Confidencialidade

- Precisão dos dados coletados

O Problema da Publicação de Dados

A Importância da Confidencialidade

- Precisão dos dados coletados
- Justiça (Fairness)

O Problema da Publicação de Dados

A Importância da Confidencialidade

- Precisão dos dados coletados
- Justiça (Fairness)
- Direitos fundamentais

O Problema da Publicação de Dados

A Importância da Confidencialidade

- Precisão dos dados coletados
- Justiça (Fairness)
- Direitos fundamentais
- Democracia

O Problema da Publicação de Dados

A Importância da Confidencialidade

- Precisão dos dados coletados
- Justiça (Fairness)
- Direitos fundamentais
- Democracia

Entretanto, garanti-la não é simples

Falha de cartórios expõe dados de ao menos 1 milhão de pais, mães e filhos

Extrato de banco de dados de certidão de nascimentos vazadas de uma associação de cartórios de São Paulo

Imagem: Reprodução

Eduardo Militão
Do UOL, em Brasília
29/10/2019 04h00

O Problema da Publicação de Dados

A Importância da Confidencialidade

- Precisão dos dados coletados
- Justiça (Fairness)
- Direitos fundamentais
- Democracia

Entretanto, garanti-la não é simples

- Dados sensíveis

Falha de cartórios expõe dados de ao menos 1 milhão de pais, mães e filhos

Extrato de banco de dados de certidão de nascimentos vazadas de uma associação de cartórios de São Paulo

Imagem: Reprodução

Eduardo Militão
Do UOL, em Brasília
29/10/2019 04h00

O Problema da Publicação de Dados

A Importância da Confidencialidade

- Precisão dos dados coletados
- Justiça (Fairness)
- Direitos fundamentais
- Democracia

Entretanto, garanti-la não é simples

- Dados sensíveis
- Quasi-identificadores

Falha de cartórios expõe dados de ao menos 1 milhão de pais, mães e filhos

Extrato de banco de dados de certidão de nascimentos vazadas de uma associação de cartórios de São Paulo

Imagem: Reprodução

Eduardo Militão
Do UOL, em Brasília
29/10/2019 04h00

Confidencialidade e Privacidade

Confidencialidade

Envolve um conjunto de regras ou uma promessa que limitam o acesso ou impõem restrições a certos tipos de informações

Confidencialidade e Privacidade

Confidencialidade

Envolve um conjunto de regras ou uma promessa que limitam o acesso ou impõem restrições a certos tipos de informações

- Usualmente utilizado no relacionamento com médicos, advogados, ou instituições financeiras

Confidencialidade e Privacidade

Confidencialidade

Envolve um conjunto de regras ou uma promessa que limitam o acesso ou impõem restrições a certos tipos de informações

- Usualmente utilizado no relacionamento com médicos, advogados, ou instituições financeiras

Privacidade

Capacidade de isolar a si mesmo ou de isolar informações sobre si mesmo e, assim, expressar-se de maneira seletiva

Confidencialidade e Privacidade

Confidencialidade

Envolve um conjunto de regras ou uma promessa que limitam o acesso ou impõem restrições a certos tipos de informações

- Usualmente utilizado no relacionamento com médicos, advogados, ou instituições financeiras

Privacidade

Capacidade de isolar a si mesmo ou de isolar informações sobre si mesmo e, assim, expressar-se de maneira seletiva

- Utilizado na Legislação mais recente sobre dados pessoais

Lei Geral de Proteção de Dados Pessoais (LGPD)

Dados pessoais são coletados:

- por governos

Lei Geral de Proteção de Dados Pessoais (LGPD)

Dados pessoais são coletados:

- por governos
- por empresas privadas

Lei Geral de Proteção de Dados Pessoais (LGPD)

Dados pessoais são coletados:

- por governos
- por empresas privadas
- para publicação

Lei Geral de Proteção de Dados Pessoais (LGPD)

Dados pessoais são coletados:

- por governos
- por empresas privadas
- para publicação
- para uso interno

Lei Geral de Proteção de Dados Pessoais (LGPD)

Dados pessoais são coletados:

- por governos
- por empresas privadas
- para publicação
- para uso interno

LGPD

- Entra em vigor em Agosto de 2020

Lei Geral de Proteção de Dados Pessoais (LGPD)

Dados pessoais são coletados:

- por governos
- por empresas privadas
- para publicação
- para uso interno

LGPD

- Entra em vigor em Agosto de 2020
- O que são dados pessoais?

Lei Geral de Proteção de Dados Pessoais (LGPD)

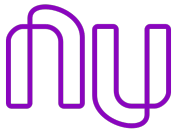
Dados pessoais são coletados:

- por governos
- por empresas privadas
- para publicação
- para uso interno

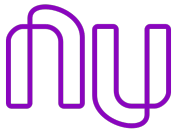
LGPD

- Entra em vigor em Agosto de 2020
- O que são dados pessoais?

Possíveis Soluções



Possíveis Soluções

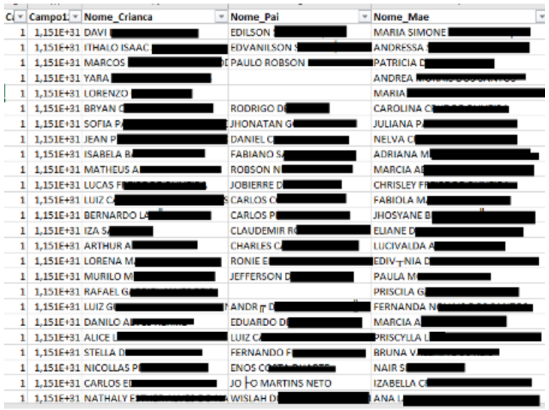


*O problema é a **forma de acesso** aos dados. Eles deveriam ser **anonimos**. A cada pessoa se **atribui um código**, um ID, e assim as pessoas que lidam com o banco de dados nao tem acesso às informações em nível pessoal. Td **criptografado**. Assim se evita invasao de privacidade, stalker, etc*

- Usuário do Twitter

Possíveis Soluções

Falha de cartórios expõe dados de ao menos 1 milhão de pais, mães e filhos



Campo1	Nome_Crianca	Nome_Pai	Nome_Mae
1	1,151E+31 DAVI [redacted]	EDILSON [redacted]	MARIA SIMONE [redacted]
1	1,151E+31 ITHALO ISAAC [redacted]	EDVANILSON [redacted]	ANDRESSA [redacted]
1	1,151E+31 MARCOS [redacted]	PAULO ROBSON [redacted]	PATRICIA [redacted]
1	1,151E+31 YARA [redacted]		ANDREA [redacted]
1	1,151E+31 LORENZO [redacted]		MARIA [redacted]
1	1,151E+31 BRYAN [redacted]	RODRIGO [redacted]	CAROLINA [redacted]
1	1,151E+31 SOFIA [redacted]	JHONATAN [redacted]	JULIANA [redacted]
1	1,151E+31 JEAN [redacted]	DANIEL [redacted]	NELVA [redacted]
1	1,151E+31 ISABELA [redacted]	FABIANO [redacted]	ADRIANA [redacted]
1	1,151E+31 MATHEUS [redacted]	ROBSON [redacted]	MARCIA [redacted]
1	1,151E+31 LUCAS [redacted]	JOBIERRE [redacted]	CHRISLEY [redacted]
1	1,151E+31 LUIZ [redacted]	CARLOS [redacted]	FABIOLA [redacted]
1	1,151E+31 BERNARDO [redacted]	CARLOS [redacted]	JHOSYANE [redacted]
1	1,151E+31 IZA [redacted]	CLAUDEMIR [redacted]	ELIANE [redacted]
1	1,151E+31 ARTHUR [redacted]	CHARLES [redacted]	LUCIVALDA [redacted]
1	1,151E+31 LORENA [redacted]	RONIE [redacted]	EDIVÂNIA [redacted]
1	1,151E+31 MURILO [redacted]	JEFFERSON [redacted]	PAULA [redacted]
1	1,151E+31 RAFAEL [redacted]		PRISCILA [redacted]
1	1,151E+31 LUIZ [redacted]	ANDRÉ [redacted]	FERNANDA [redacted]
1	1,151E+31 DANILO [redacted]	EDUARDO [redacted]	MARCIA [redacted]
1	1,151E+31 ALICE [redacted]	LUIZ [redacted]	PRISCYLLA [redacted]
1	1,151E+31 STELLA [redacted]	FERNANDO [redacted]	BRUNA [redacted]
1	1,151E+31 NICOLLAS [redacted]	ENOS [redacted]	NAIR [redacted]
1	1,151E+31 CARLOS [redacted]	JOÃO MARTINS NETO	IZABELLA [redacted]
1	1,151E+31 NATHALY [redacted]	WISLAH [redacted]	ANA [redacted]

Extrato de banco de dados de certidão de nascimentos vazadas de uma associação de cartórios de São Paulo

Imagem: Reprodução

Eduardo Militão
Do UOL, em Brasília
29/10/2019 04h00

Possíveis Soluções

Falha de cartórios expõe dados de ao menos 1 milhão de pais, mães e filhos

Campo1	Nome_Crianca	Nome_Pai	Nome_Mae
1	1,151E+31 DAVI [redacted]	EDILSON [redacted]	MARIA SIMONE [redacted]
1	1,151E+31 ITHALO ISAAC [redacted]	EDVANILSON [redacted]	ANDRESSA [redacted]
1	1,151E+31 MARCOS [redacted]	PAULO ROBSON [redacted]	PATRICIA [redacted]
1	1,151E+31 YARA [redacted]		ANDREA [redacted]
1	1,151E+31 LORENZO [redacted]		MARIA [redacted]
1	1,151E+31 BRYAN [redacted]	RODRIGO [redacted]	CAROLINA [redacted]
1	1,151E+31 SOFIA [redacted]	JHONATAN [redacted]	JULIANA [redacted]
1	1,151E+31 JEAN [redacted]	DANIEL [redacted]	NELVA [redacted]
1	1,151E+31 ISABELA [redacted]	FABIANO [redacted]	ADRIANA [redacted]
1	1,151E+31 MATHEUS [redacted]	ROBSON [redacted]	MARCIA [redacted]
1	1,151E+31 LUCAS [redacted]	JOBIERRE [redacted]	CHRISLEY [redacted]
1	1,151E+31 LUIZ [redacted]	CARLOS [redacted]	FABIOLA [redacted]
1	1,151E+31 BERNARDO [redacted]	CARLOS [redacted]	JHOSYANE [redacted]
1	1,151E+31 IZA [redacted]	CLAUDEMIR [redacted]	ELIANE [redacted]
1	1,151E+31 ARTHUR [redacted]	CHARLES [redacted]	LUCIVALDA [redacted]
1	1,151E+31 LORENA [redacted]	RONIE [redacted]	EDIVÂNIA [redacted]
1	1,151E+31 MURILO [redacted]	JEFFERSON [redacted]	PAULA [redacted]
1	1,151E+31 RAFAEL [redacted]		PRISCILA [redacted]
1	1,151E+31 LUIZ [redacted]	ANDRÉ [redacted]	FERNANDA [redacted]
1	1,151E+31 DANILO [redacted]	EDUARDO [redacted]	MARCIA [redacted]
1	1,151E+31 ALICE [redacted]	LUIZ [redacted]	PRISCYLLA [redacted]
1	1,151E+31 STELLA [redacted]	FERNANDO [redacted]	BRUNA [redacted]
1	1,151E+31 NICOLLAS [redacted]	ENOS [redacted]	NAIR [redacted]
1	1,151E+31 CARLOS [redacted]	JOÃO MARTINS NETO	IZABELLA [redacted]
1	1,151E+31 NATHALY [redacted]	WISLAH [redacted]	ANA [redacted]

Extrato de banco de dados de certidão de nascimentos vazadas de uma associação de cartórios de São Paulo

Imagem: Reprodução

Eduardo Militão
Do UOL, em Brasília
29/10/2019 04h00

Controle de acesso

Possíveis Soluções

Falha de cartórios expõe dados de ao menos 1 milhão de pais, mães e filhos

Campo1	Nome_Crianca	Nome_Pai	Nome_Mae
1	1,151E+31 DAVI [redacted]	EDILSON [redacted]	MARIA SIMONE [redacted]
1	1,151E+31 ITHALO ISAAC [redacted]	EDVANILSON [redacted]	ANDRESSA [redacted]
1	1,151E+31 MARCOS [redacted]	PAULO ROBSON [redacted]	PATRICIA [redacted]
1	1,151E+31 YARA [redacted]		ANDREA [redacted]
1	1,151E+31 LORENZO [redacted]		MARIA [redacted]
1	1,151E+31 BRYAN [redacted]	RODRIGO [redacted]	CAROLINA [redacted]
1	1,151E+31 SOFIA [redacted]	JHONATAN [redacted]	JULIANA [redacted]
1	1,151E+31 JEAN [redacted]	DANIEL [redacted]	NELVA [redacted]
1	1,151E+31 ISABELA [redacted]	FABIANO [redacted]	ADRIANA [redacted]
1	1,151E+31 MATHEUS [redacted]	ROBSON [redacted]	MARCIA [redacted]
1	1,151E+31 LUCAS [redacted]	JOBIERRE [redacted]	CHRISLEY [redacted]
1	1,151E+31 LUIZ [redacted]	CARLOS [redacted]	FABIOLA [redacted]
1	1,151E+31 BERNARDO [redacted]	CARLOS [redacted]	JHOSYANE [redacted]
1	1,151E+31 IZA [redacted]	CLAUDEMIR [redacted]	ELIANE [redacted]
1	1,151E+31 ARTHUR [redacted]	CHARLES [redacted]	LUCIVALDA [redacted]
1	1,151E+31 LORENA [redacted]	RONIE [redacted]	EDIVÂNIA [redacted]
1	1,151E+31 MURILO [redacted]	JEFFERSON [redacted]	PAULA [redacted]
1	1,151E+31 RAFAEL [redacted]		PRISCILA [redacted]
1	1,151E+31 LUIZ [redacted]	ANDR [redacted]	FERNANDA [redacted]
1	1,151E+31 DANILO [redacted]	EDUARDO [redacted]	MARCIA [redacted]
1	1,151E+31 ALICE [redacted]	LUIZ [redacted]	PRISCYLLA [redacted]
1	1,151E+31 STELLA [redacted]	FERNANDO [redacted]	BRUNA [redacted]
1	1,151E+31 NICOLLAS [redacted]	ENOS [redacted]	NAIR [redacted]
1	1,151E+31 CARLOS [redacted]	JO [redacted]O MARTINS NETO	IZABELLA [redacted]
1	1,151E+31 NATHALY [redacted]	WISLAH [redacted]	ANA [redacted]

Extrato de banco de dados de certidão de nascimentos vazadas de uma associação de cartórios de São Paulo

Imagem: Reprodução

Eduardo Militão
Do UOL, em Brasília
29/10/2019 04h00

Controle de acesso

- Evita acesso não autorizado

Possíveis Soluções

Falha de cartórios expõe dados de ao menos 1 milhão de pais, mães e filhos

Campo1	Nome_Crianca	Nome_Pai	Nome_Mae
1	1,151E+31 DAVI [redacted]	EDILSON [redacted]	MARIA SIMONE [redacted]
1	1,151E+31 ITHALO ISAAC [redacted]	EDVANILSON [redacted]	ANDRESSA [redacted]
1	1,151E+31 MARCOS [redacted]	PAULO ROBSON [redacted]	PATRICIA [redacted]
1	1,151E+31 YARA [redacted]		ANDREA [redacted]
1	1,151E+31 LORENZO [redacted]		MARIA [redacted]
1	1,151E+31 BRYAN [redacted]	RODRIGO [redacted]	CAROLINA [redacted]
1	1,151E+31 SOFIA [redacted]	JHONATAN [redacted]	JULIANA [redacted]
1	1,151E+31 JEAN [redacted]	DANIEL [redacted]	NELVA [redacted]
1	1,151E+31 ISABELA [redacted]	FABIANO [redacted]	ADRIANA [redacted]
1	1,151E+31 MATHEUS [redacted]	ROBSON [redacted]	MARCIA [redacted]
1	1,151E+31 LUCAS [redacted]	JOBIERRE [redacted]	CHRISLEY [redacted]
1	1,151E+31 LUIZ [redacted]	CARLOS [redacted]	FABIOLA [redacted]
1	1,151E+31 BERNARDO [redacted]	CARLOS [redacted]	JHOSYANE [redacted]
1	1,151E+31 IZA [redacted]	CLAUDEMIR [redacted]	ELIANE [redacted]
1	1,151E+31 ARTHUR [redacted]	CHARLES [redacted]	LUCIVALDA [redacted]
1	1,151E+31 LORENA [redacted]	RONIE [redacted]	EDIVÂNIA [redacted]
1	1,151E+31 MURILO [redacted]	JEFFERSON [redacted]	PAULA [redacted]
1	1,151E+31 RAFAEL [redacted]		PRISCILA [redacted]
1	1,151E+31 LUIZ [redacted]	ANDR [redacted]	FERNANDA [redacted]
1	1,151E+31 DANILO [redacted]	EDUARDO [redacted]	MARCIA [redacted]
1	1,151E+31 ALICE [redacted]	LUIZ [redacted]	PRISCYLLA [redacted]
1	1,151E+31 STELLA [redacted]	FERNANDO [redacted]	BRUNA [redacted]
1	1,151E+31 NICOLLAS [redacted]	ENOS [redacted]	NAIR [redacted]
1	1,151E+31 CARLOS [redacted]	JO [redacted]O MARTINS NETO	IZABELLA [redacted]
1	1,151E+31 NATHALY [redacted]	WISLAH [redacted]	ANA [redacted]

Extrato de banco de dados de certidão de nascimentos vazadas de uma associação de cartórios de São Paulo

Imagem: Reprodução

Eduardo Militão
Do UOL, em Brasília
29/10/2019 04h00

Controle de acesso

- Evita acesso não autorizado
- Evita vazamento de informações

Possíveis Soluções

Falha de cartórios expõe dados de ao menos 1 milhão de pais, mães e filhos

Campo1	Nome_Crianca	Nome_Pai	Nome_Mae
1	1,151E+31 DAVI [redacted]	EDILSON [redacted]	MARIA SIMONE [redacted]
1	1,151E+31 ITHALO ISAAC [redacted]	EDVANILSON [redacted]	ANDRESSA [redacted]
1	1,151E+31 MARCOS [redacted]	PAULO ROBSON [redacted]	PATRICIA [redacted]
1	1,151E+31 YARA [redacted]		ANDREA [redacted]
1	1,151E+31 LORENZO [redacted]		MARIA [redacted]
1	1,151E+31 BRYAN [redacted]	RODRIGO [redacted]	CAROLINA [redacted]
1	1,151E+31 SOFIA [redacted]	JHONATAN [redacted]	JULIANA [redacted]
1	1,151E+31 JEAN [redacted]	DANIEL [redacted]	NELVA [redacted]
1	1,151E+31 ISABELA [redacted]	FABIANO [redacted]	ADRIANA [redacted]
1	1,151E+31 MATHEUS [redacted]	ROBSON [redacted]	MARCIA [redacted]
1	1,151E+31 LUCAS [redacted]	JOBIERRE [redacted]	CHRISLEY [redacted]
1	1,151E+31 LUIZ [redacted]	CARLOS [redacted]	FABIOLA [redacted]
1	1,151E+31 BERNARDO [redacted]	CARLOS [redacted]	JHOSYANE [redacted]
1	1,151E+31 IZA [redacted]	CLAUDEMIR [redacted]	ELIANE [redacted]
1	1,151E+31 ARTHUR [redacted]	CHARLES [redacted]	LUCIVALDA [redacted]
1	1,151E+31 LORENA [redacted]	RONIE [redacted]	EDIVÂNIA [redacted]
1	1,151E+31 MURILO [redacted]	JEFFERSON [redacted]	PAULA [redacted]
1	1,151E+31 RAFAEL [redacted]		PRISCILA [redacted]
1	1,151E+31 LUIZ [redacted]	ANDR [redacted]	FERNANDA [redacted]
1	1,151E+31 DANILO [redacted]	EDUARDO [redacted]	MARCIA [redacted]
1	1,151E+31 ALICE [redacted]	LUIZ [redacted]	PRISCYLLA [redacted]
1	1,151E+31 STELLA [redacted]	FERNANDO [redacted]	BRUNA [redacted]
1	1,151E+31 NICOLLAS [redacted]	ENOS [redacted]	NAIR [redacted]
1	1,151E+31 CARLOS [redacted]	JO [redacted]O MARTINS NETO	IZABELLA [redacted]
1	1,151E+31 NATHALY [redacted]	WISLAH [redacted]	ANA [redacted]

Extrato de banco de dados de certidão de nascimentos vazadas de uma associação de cartórios de São Paulo

Imagem: Reprodução

Eduardo Militão
Do UOL, em Brasília
29/10/2019 04h00

Controle de acesso e Criptografia

- Evitam acesso não autorizado
- Evitam vazamento de informações

Possíveis Soluções

Falha de cartórios expõe dados de ao menos 1 milhão de pais, mães e filhos

Campo1	Nome_Crianca	Nome_Pai	Nome_Mae
1	1,151E+31 DAVI [redacted]	EDILSON [redacted]	MARIA SIMONE [redacted]
1	1,151E+31 ITHALO ISAAC [redacted]	EDVANILSON [redacted]	ANDRESSA [redacted]
1	1,151E+31 MARCOS [redacted]	PAULO ROBSON [redacted]	PATRICIA [redacted]
1	1,151E+31 YARA [redacted]		ANDREA [redacted]
1	1,151E+31 LORENZO [redacted]		MARIA [redacted]
1	1,151E+31 BRYAN [redacted]	RODRIGO [redacted]	CAROLINA [redacted]
1	1,151E+31 SOFIA [redacted]	JHONATAN [redacted]	JULIANA [redacted]
1	1,151E+31 JEAN [redacted]	DANIEL [redacted]	NELVA [redacted]
1	1,151E+31 ISABELA [redacted]	FABIANO [redacted]	ADRIANA [redacted]
1	1,151E+31 MATHEUS [redacted]	ROBSON [redacted]	MARCIA [redacted]
1	1,151E+31 LUCAS [redacted]	JOBIERRE [redacted]	CHRISLEY [redacted]
1	1,151E+31 LUIZ [redacted]	CARLOS [redacted]	FABIOLA [redacted]
1	1,151E+31 BERNARDO [redacted]	CARLOS [redacted]	JHOSYANE [redacted]
1	1,151E+31 IZA [redacted]	CLAUDEMIR [redacted]	ELIANE [redacted]
1	1,151E+31 ARTHUR [redacted]	CHARLES [redacted]	LUCIVALDA [redacted]
1	1,151E+31 LORENA [redacted]	RONIE [redacted]	EDIVÂNIA [redacted]
1	1,151E+31 MURILO [redacted]	JEFFERSON [redacted]	PAULA [redacted]
1	1,151E+31 RAFAEL [redacted]		PRISCILA [redacted]
1	1,151E+31 LUIZ [redacted]	ANDR [redacted]	FERNANDA [redacted]
1	1,151E+31 DANILO [redacted]	EDUARDO [redacted]	MARCIA [redacted]
1	1,151E+31 ALICE [redacted]	LUIZ [redacted]	PRISCYLLA [redacted]
1	1,151E+31 STELLA [redacted]	FERNANDO [redacted]	BRUNA [redacted]
1	1,151E+31 NICOLLAS [redacted]	ENOS [redacted]	NAIR [redacted]
1	1,151E+31 CARLOS [redacted]	JO [redacted]O MARTINS NETO	IZABELLA [redacted]
1	1,151E+31 NATHALY [redacted]	WISLAH [redacted]	ANA [redacted]

Extrato de banco de dados de certidão de nascimentos vazadas de uma associação de cartórios de São Paulo

Imagem: Reprodução

Eduardo Militão
Do UOL, em Brasília
29/10/2019 04h00

Controle de acesso e Criptografia

- Evitam acesso não autorizado
- Evitam vazamento de informações
- Não protegem contra inferência quando os dados são acessíveis

Possíveis Soluções

Falha de cartórios expõe dados de ao menos 1 milhão de pais, mães e filhos

Campo1	Nome_Crianca	Nome_Pai	Nome_Mae
1	1,151E+31 DAVI [redacted]	EDILSON [redacted]	MARIA SIMONE [redacted]
1	1,151E+31 ITHALO ISAAC [redacted]	EDVANILSON [redacted]	ANDRESSA [redacted]
1	1,151E+31 MARCOS [redacted]	PAULO ROBSON [redacted]	PATRICIA [redacted]
1	1,151E+31 YARA [redacted]		ANDREA [redacted]
1	1,151E+31 LORENZO [redacted]		MARIA [redacted]
1	1,151E+31 BRYAN [redacted]	RODRIGO [redacted]	CAROLINA [redacted]
1	1,151E+31 SOFIA [redacted]	JHONATAN [redacted]	JULIANA [redacted]
1	1,151E+31 JEAN [redacted]	DANIEL [redacted]	NELVA [redacted]
1	1,151E+31 ISABELA [redacted]	FABIANO [redacted]	ADRIANA [redacted]
1	1,151E+31 MATHEUS [redacted]	ROBSON [redacted]	MARCIA [redacted]
1	1,151E+31 LUCAS [redacted]	JOBIERRE [redacted]	CHRISLEY [redacted]
1	1,151E+31 LUIZ [redacted]	CARLOS [redacted]	FABIOLA [redacted]
1	1,151E+31 BERNARDO [redacted]	CARLOS [redacted]	JHOSYANE [redacted]
1	1,151E+31 IZA [redacted]	CLAUDEMIR [redacted]	ELIANE [redacted]
1	1,151E+31 ARTHUR [redacted]	CHARLES [redacted]	LUCIVALDA [redacted]
1	1,151E+31 LORENA [redacted]	RONIE [redacted]	EDIVÂNIA [redacted]
1	1,151E+31 MURILO [redacted]	JEFFERSON [redacted]	PAULA [redacted]
1	1,151E+31 RAFAEL [redacted]		PRISCILA [redacted]
1	1,151E+31 LUIZ [redacted]	ANDR [redacted]	FERNANDA [redacted]
1	1,151E+31 DANILO [redacted]	EDUARDO [redacted]	MARCIA [redacted]
1	1,151E+31 ALICE [redacted]	LUIZ [redacted]	PRISCYLLA [redacted]
1	1,151E+31 STELLA [redacted]	FERNANDO [redacted]	BRUNA [redacted]
1	1,151E+31 NICOLLAS [redacted]	ENOS [redacted]	NAIR [redacted]
1	1,151E+31 CARLOS [redacted]	JO [redacted]O MARTINS NETO	IZABELLA [redacted]
1	1,151E+31 NATHALY [redacted]	WISLAH [redacted]	ANA [redacted]

Extrato de banco de dados de certidão de nascimentos vazadas de uma associação de cartórios de São Paulo

Imagem: Reprodução

Eduardo Militão
Do UOL, em Brasília
29/10/2019 04h00

- Controle de acesso e Criptografia
  - Evitam acesso não autorizado
  - Evitam vazamento de informações
  - Não protegem contra inferência quando os dados são acessíveis
- Anonimização
  - Diversos métodos propostos

Possíveis Soluções

Falha de cartórios expõe dados de ao menos 1 milhão de pais, mães e filhos

Campo1	Nome_Crianca	Nome_Pai	Nome_Mae
1	1,151E+31 DAVI [redacted]	EDILSON [redacted]	MARIA SIMONE [redacted]
1	1,151E+31 ITHALO ISAAC [redacted]	EDVANILSON [redacted]	ANDRESSA [redacted]
1	1,151E+31 MARCOS [redacted]	PAULO ROBSON [redacted]	PATRICIA [redacted]
1	1,151E+31 YARA [redacted]		ANDREA [redacted]
1	1,151E+31 LORENZO [redacted]		MARIA [redacted]
1	1,151E+31 BRYAN [redacted]	RODRIGO [redacted]	CAROLINA [redacted]
1	1,151E+31 SOFIA [redacted]	JHONATAN [redacted]	JULIANA [redacted]
1	1,151E+31 JEAN [redacted]	DANIEL [redacted]	NELVA [redacted]
1	1,151E+31 ISABELA [redacted]	FABIANO [redacted]	ADRIANA [redacted]
1	1,151E+31 MATHEUS [redacted]	ROBSON [redacted]	MARCIA [redacted]
1	1,151E+31 LUCAS [redacted]	JOBIERRE [redacted]	CHRISLEY [redacted]
1	1,151E+31 LUIZ [redacted]	CARLOS [redacted]	FABIOLA [redacted]
1	1,151E+31 BERNARDO [redacted]	CARLOS [redacted]	JHOSYANE [redacted]
1	1,151E+31 IZA [redacted]	CLAUDEMIR [redacted]	ELIANE [redacted]
1	1,151E+31 ARTHUR [redacted]	CHARLES [redacted]	LUCIVALDA [redacted]
1	1,151E+31 LORENA [redacted]	RONIE [redacted]	EDIVÂNIA [redacted]
1	1,151E+31 MURILO [redacted]	JEFFERSON [redacted]	PAULA [redacted]
1	1,151E+31 RAFAEL [redacted]		PRISCILA [redacted]
1	1,151E+31 LUIZ [redacted]	ANDRÉ [redacted]	FERNANDA [redacted]
1	1,151E+31 DANILO [redacted]	EDUARDO [redacted]	MARCIA [redacted]
1	1,151E+31 ALICE [redacted]	LUIZ [redacted]	PRISCYLLA [redacted]
1	1,151E+31 STELLA [redacted]	FERNANDO [redacted]	BRUNA [redacted]
1	1,151E+31 NICOLLAS [redacted]	ENOS [redacted]	NAIR [redacted]
1	1,151E+31 CARLOS [redacted]	JOÃO MARTINS NETO	IZABELLA [redacted]
1	1,151E+31 NATHALY [redacted]	WISLAH [redacted]	ANA [redacted]

Extrato de banco de dados de certidão de nascimentos vazadas de uma associação de cartórios de São Paulo

Imagem: Reprodução

Eduardo Militão
Do UOL, em Brasília
29/10/2019 04h00

Controle de acesso e Criptografia

- Evitam acesso não autorizado
- Evitam vazamento de informações
- Não protegem contra inferência quando os dados são acessíveis

Anonimização

- Diversos métodos propostos
- Diversos vetores de ataque já demonstrados

Anonimização

Desidentificação e Pseudonimização

Desidentificação ou Anonimização "inocente"

Definição

Uma base de dados é dita anônima se dela foram retirados os identificadores diretos dos indivíduos

Desidentificação ou Anonimização "inocente"

Definição

Uma base de dados é dita anônima se dela foram retirados os identificadores diretos dos indivíduos

Exemplos

Nome, números únicos de identificação (CPF, RG), endereço, etc

Desidentificação ou Anonimização "inocente"

Definição

Uma base de dados é dita anônima se dela foram retirados os identificadores diretos dos indivíduos

Exemplos

Nome, números únicos de identificação (CPF, RG), endereço, etc

Entretanto, a definição de identificadores diretos é muito flexível

Exemplo de Desidentificação

	Nome	Idade	Condição
1	Jon Snow	30	Resfriado
2	Jamie Lannister	39	Mão amputada
3	Arya Stark	16	Dor de estômago
4	Bran Stark	14	Paraplegia
5	Eddad Stark	32	Dor de cabeça
6	Ramsay Bolton	32	Psicopatia
7	Daenerys Targaryen	25	Mania de grandeza

Catuscia Palamidessi, Kostas Chatzikokolakis

Exemplo de Desidentificação

	Nome	Idade	Condição
1	*	30	Resfriado
2	*	39	Mão amputada
3	*	16	Dor de estômago
4	*	14	Paraplegia
5	*	32	Dor de cabeça
6	*	32	Psicopatia
7	*	25	Mania de grandeza

Quasi-identificadores

Definição

Atributos que podem ser vinculados a informações externas para identificar indivíduos unicamente

Quasi-identificadores

Definição

Atributos que podem ser vinculados a informações externas para identificar indivíduos unicamente

Observação

Definir o conjunto de quasi-identificadores para uma dada publicação ainda é uma questão em aberto

Sweeney's Linkage Attack (1998)

DB1: Contém dados sensíveis (anonimizada)

Sweeney's Linkage Attack (1998)

DB1: Contém dados sensíveis (anonimizada)

DB2: Coleção pública de dados não sensíveis

Sweeney's Linkage Attack (1998)

DB1: Contém dados sensíveis (anonimizada)

DB2: Coleção pública de dados não sensíveis

Informações auxiliares

Sweeney's Linkage Attack (1998)

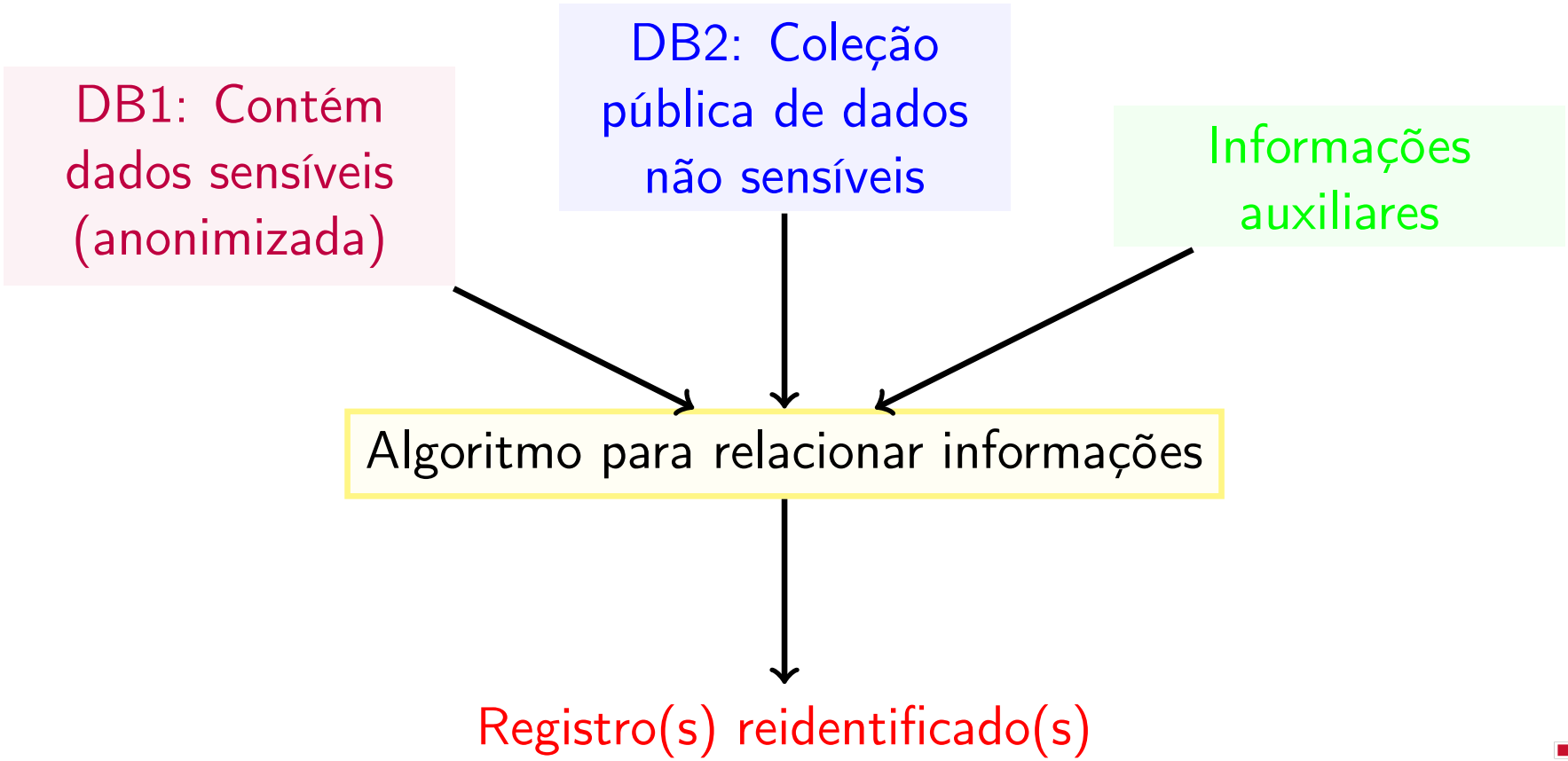
DB1: Contém dados sensíveis (anonimizada)

DB2: Coleção pública de dados não sensíveis

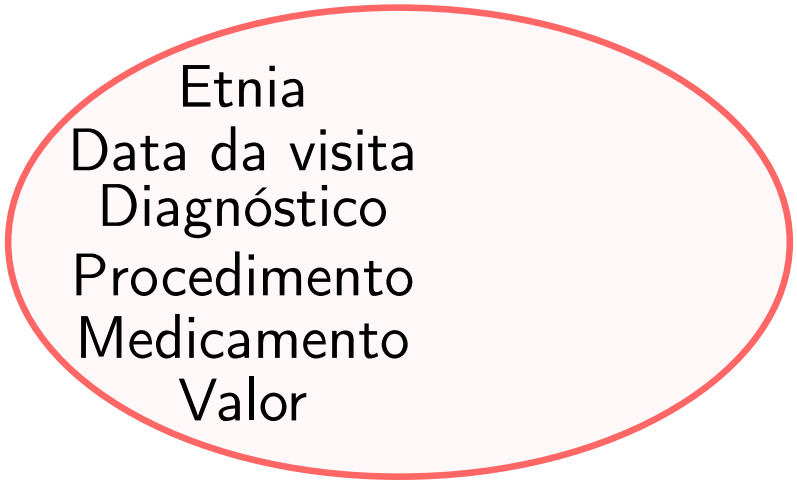
Informações auxiliares

Algoritmo para relacionar informações

Sweeney's Linkage Attack (1998)

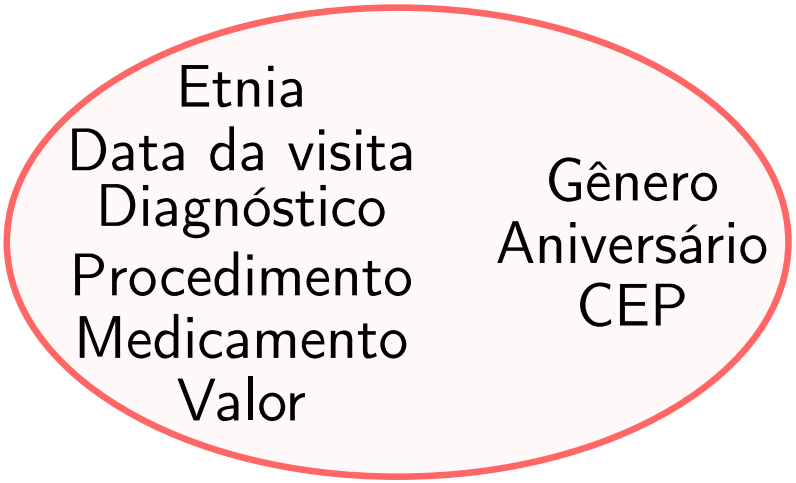


Sweeney's Linkage Attack na Prática



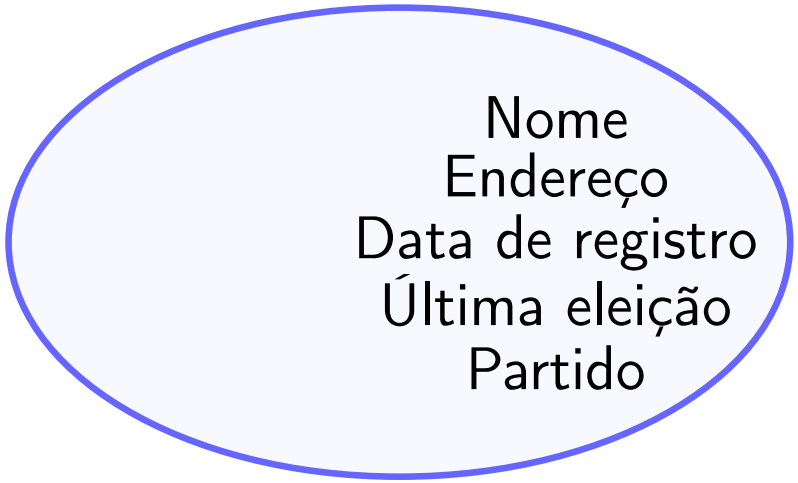
Base 1: Dados médicos

Sweeney's Linkage Attack na Prática



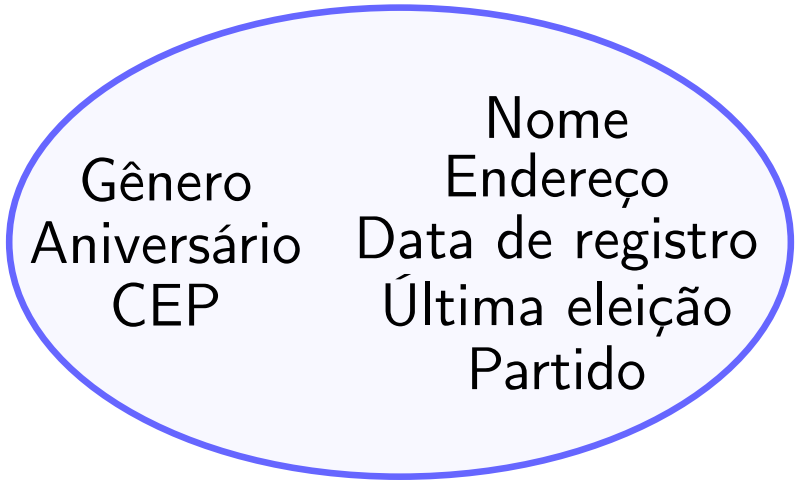
Base 1: Dados médicos

Sweeney's Linkage Attack na Prática



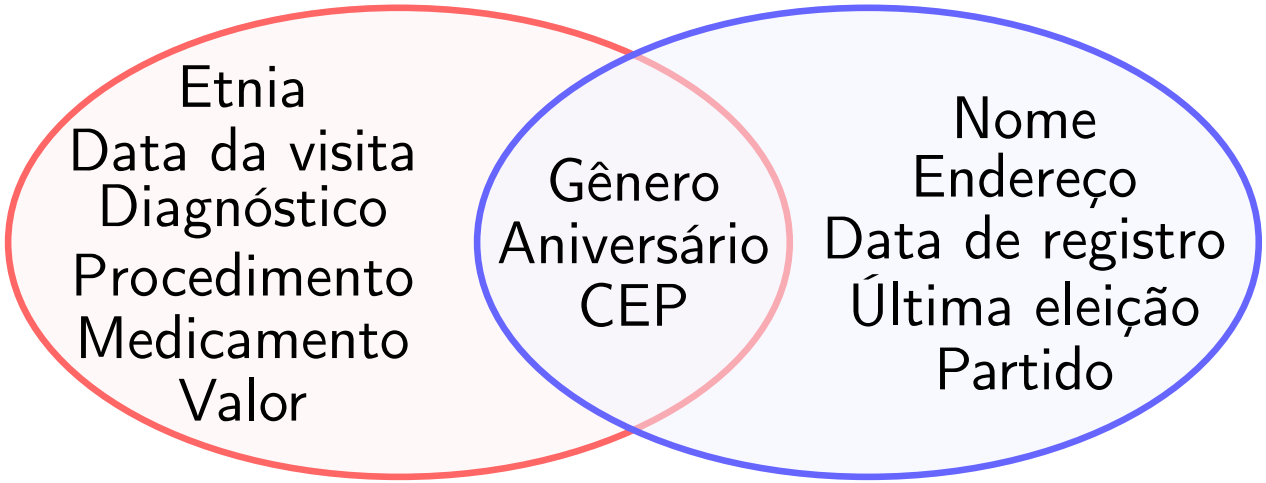
Base 2: Lista de eleitores

Sweeney's Linkage Attack na Prática



Base 2: Lista de eleitores

Sweeney's Linkage Attack na Prática



Base 1: Dados médicos

Base 2: Lista de eleitores

Sweeney's Linkage Attack na Prática

Conclusões

- Com apenas **gênero**, **data de aniversário**, e **CEP**, Sweeney foi capaz de reidentificar o governador de Massachusetts

Sweeney's Linkage Attack na Prática

Conclusões

- Com apenas **gênero**, **data de aniversário**, e **CEP**, Sweeney foi capaz de reidentificar o governador de Massachusetts
- O mesmo foi possível para **87% da população** dos EUA, também unicamente identificável pelos mesmos atributos

Pseudonimização

Definição

Uma base de dados é dita pseudonimizada se dela foram retirados os identificadores diretos dos indivíduos em substituição por um código único para cada indivíduo

Pseudonimização

Definição

Uma base de dados é dita pseudonimizada se dela foram retirados os identificadores diretos dos indivíduos em substituição por um código único para cada indivíduo

Entretanto...

- Apresenta os mesmos problemas que a Desidentificação

Pseudonimização

Definição

Uma base de dados é dita pseudonimizada se dela foram retirados os identificadores diretos dos indivíduos em substituição por um código único para cada indivíduo

Entretanto...

- Apresenta os mesmos problemas que a Desidentificação
- Publicação de dados longitudinais no tempo

Pseudonimização

Definição

Uma base de dados é dita pseudonimizada se dela foram retirados os identificadores diretos dos indivíduos em substituição por um código único para cada indivíduo

Entretanto...

- Apresenta os mesmos problemas que a Desidentificação
- Publicação de dados longitudinais no tempo
  - Pseudonimização é ainda pior do que Desidentificação

Pseudonimização

Definição

Uma base de dados é dita pseudonimizada se dela foram retirados os identificadores diretos dos indivíduos em substituição por um código único para cada indivíduo

Entretanto...

- Apresenta os mesmos problemas que a Desidentificação
- Publicação de dados longitudinais no tempo
  - Pseudonimização é ainda pior do que Desidentificação
  - Facilita a reidentificação ao longo do tempo

Exemplo de Pseudonimização

Exemplo de Pseudonimização

- Divulgou 20 milhões de pesquisas realizadas por mais de 650 mil usuários ao longo de três meses

Exemplo de Pseudonimização

- Divulgou 20 milhões de pesquisas realizadas por mais de 650 mil usuários ao longo de três meses
- Objetivo era fomentar pesquisas científicas

Exemplo de Pseudonimização

- Divulgou 20 milhões de pesquisas realizadas por mais de 650 mil usuários ao longo de três meses
- Objetivo era fomentar pesquisas científicas
- Os dados foram liberados para toda a Internet

Exemplo de Pseudonimização

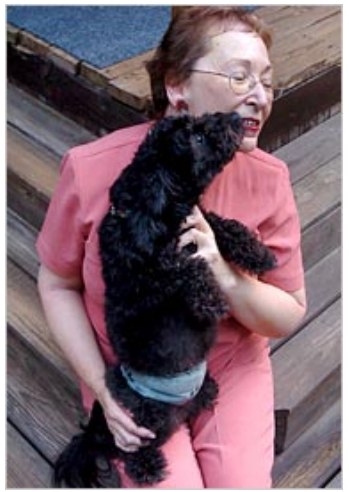
- Divulgou 20 milhões de pesquisas realizadas por mais de 650 mil usuários ao longo de três meses
- Objetivo era fomentar pesquisas científicas
- Os dados foram liberados para toda a Internet
- *The New York Times* cruzou as pesquisas de um usuário com listas telefônicas

Exemplo de Pseudonimização

- Divulgou 20 milhões de pesquisas realizadas por mais de 650 mil usuários ao longo de três meses
- Objetivo era fomentar pesquisas científicas
- Os dados foram liberados para toda a Internet
- *The New York Times* cruzou as pesquisas de um usuário com listas telefônicas
  - "60 anos de idade", "homens solteiros", "cão que urina em tudo", "paisagistas em Lilburn, GA"

Exemplo de Pseudonimização

- Divulgou 20 milhões de pesquisas realizadas por mais de 650 mil usuários ao longo de três meses
- Objetivo era fomentar pesquisas científicas
- Os dados foram liberados para toda a Internet
- *The New York Times* cruzou as pesquisas de um usuário com listas telefônicas
  - "60 anos de idade", "homens solteiros", "cão que urina em tudo", "paisagistas em Lilburn, GA"
  - Reidentificou Thelma Arnold, viúva de 62 anos, com três cães, de Lilburn, Geórgia



Anonimização

Métodos Determinísticos

k-Anonimização (Sweeney & Samarati, 1998)

Definição

Uma base de dados é dita k-Anônima se cada registro for indistinguível de $k-1$ outros registros, considerando-se os quase-identificadores, através de:

k-Anonimização (Sweeney & Samarati, 1998)

Definição

Uma base de dados é dita k-Anônima se cada registro for indistinguível de $k-1$ outros registros, considerando-se os quase-identificadores, através de:

- generalização de atributos

k-Anonimização (Sweeney & Samarati, 1998)

Definição

Uma base de dados é dita k-Anônima se cada registro for indistinguível de $k-1$ outros registros, considerando-se os quase-identificadores, através de:

- generalização de atributos
- supressão de atributos

k-Anonimização (Sweeney & Samarati, 1998)

Definição

Uma base de dados é dita k-Anônima se cada registro for indistinguível de $k-1$ outros registros, considerando-se os quase-identificadores, através de:

- generalização de atributos
- supressão de atributos
- adição de registros sintéticos

Exemplo de 4-Anonimização

	Não Sensível			Sensível
	CEP	Idade	País	Condição
1	13053	28	Rússia	Cardíaco
2	13068	29	EUA	Cardíaco
3	13068	21	Japão	Virose
4	13053	23	EUA	Virose
5	14853	50	Índia	Câncer
6	14853	55	Rússia	Cardíaco
7	14850	47	EUA	Virose
8	14850	49	EUA	Virose
9	13053	31	EUA	Câncer
10	13053	37	Índia	Câncer
11	13068	36	Japão	Câncer
12	13068	35	EUA	Câncer

	Não Sensível			Sensível
	CEP	Idade	País	Condição
1	130**	< 30	*	Cardíaco
2	130**	< 30	*	Cardíaco
3	130**	< 30	*	Virose
4	130**	< 30	*	Virose
5	1485*	≥ 40	*	Câncer
6	1485*	≥ 40	*	Cardíaco
7	1485*	≥ 40	*	Virose
8	1485*	≥ 40	*	Virose
9	130**	3*	*	Câncer
10	130**	3*	*	Câncer
11	130**	3*	*	Câncer
12	130**	3*	*	Câncer

Exemplo de 4-Anonimização

	Não Sensível			Sensível
	CEP	Idade	País	Condição
1	130**	< 30	*	Cardíaco
2	130**	< 30	*	Cardíaco
3	130**	< 30	*	Virose
4	130**	< 30	*	Virose
5	1485*	≥ 40	*	Câncer
6	1485*	≥ 40	*	Cardíaco
7	1485*	≥ 40	*	Virose
8	1485*	≥ 40	*	Virose
9	130**	3*	*	Câncer
10	130**	3*	*	Câncer
11	130**	3*	*	Câncer
12	130**	3*	*	Câncer

Exemplo de 4-Anonimização

	Não Sensível			Sensível
	CEP	Idade	País	Condição
1	130**	< 30	*	Cardíaco
2	130**	< 30	*	Cardíaco
3	130**	< 30	*	Virose
4	130**	< 30	*	Virose
5	1485*	≥ 40	*	Câncer
6	1485*	≥ 40	*	Cardíaco
7	1485*	≥ 40	*	Virose
8	1485*	≥ 40	*	Virose
9	130**	3*	*	Câncer
10	130**	3*	*	Câncer
11	130**	3*	*	Câncer
12	130**	3*	*	Câncer

Exemplo de 4-Anonimização

	Não Sensível			Sensível
	CEP	Idade	País	Condição
1	130**	< 30	*	Cardíaco
2	130**	< 30	*	Cardíaco
3	130**	< 30	*	Virose
4	130**	< 30	*	Virose
5	1485*	≥ 40	*	Câncer
6	1485*	≥ 40	*	Cardíaco
7	1485*	≥ 40	*	Virose
8	1485*	≥ 40	*	Virose
9	130**	3*	*	Câncer
10	130**	3*	*	Câncer
11	130**	3*	*	Câncer
12	130**	3*	*	Câncer

ℓ-Diversidade (Kifer et al., 2007)

Definição

Uma base de dados é dita ℓ-Diversa se cada agrupamento de registros apresentar uma diversidade de, ao menos, ℓ atributos sensíveis

Exemplo de 3-Diversidade

	Não Sensível			Sensível
	CEP	Idade	Gênero	Condição
1	130**	≤ 50	*	Câncer
2	130**	≤ 50	*	Câncer
3	130**	≤ 50	*	Hemofilia
4	130**	≤ 50	*	Virose
5	122**	> 50	*	Hemofilia
6	122**	> 50	*	Câncer
7	122**	> 50	*	Virose
8	122**	> 50	*	Virose
9	130**	≤ 50	*	Câncer
10	130**	≤ 50	*	Câncer
11	130**	≤ 50	*	Hemofilia
12	130**	≤ 50	*	Virose

Exemplo de 3-Diversidade

	Não Sensível			Sensível
	CEP	Idade	Gênero	Condição
1	130**	≤ 50	*	Câncer
2	130**	≤ 50	*	Câncer
3	130**	≤ 50	*	Hemofilia
4	130**	≤ 50	*	Virose
5	122**	> 50	*	Hemofilia
6	122**	> 50	*	Câncer
7	122**	> 50	*	Virose
8	122**	> 50	*	Virose
9	130**	≤ 50	*	Câncer
10	130**	≤ 50	*	Câncer
11	130**	≤ 50	*	Hemofilia
12	130**	≤ 50	*	Virose

Exemplo de 3-Diversidade

	Não Sensível			Sensível
	CEP	Idade	Gênero	Condição
1	130**	≤ 50	*	Câncer
2	130**	≤ 50	*	Câncer
3	130**	≤ 50	*	Hemofilia
4	130**	≤ 50	*	Virose
5	122**	> 50	*	Hemofilia
6	122**	> 50	*	Câncer
7	122**	> 50	*	Virose
8	122**	> 50	*	Virose
9	130**	≤ 50	*	Câncer
10	130**	≤ 50	*	Câncer
11	130**	≤ 50	*	Hemofilia
12	130**	≤ 50	*	Virose

Exemplo de 3-Diversidade

	Não Sensível			Sensível
	CEP	Idade	Gênero	Condição
1	130**	≤ 50	*	Câncer
2	130**	≤ 50	*	Câncer
3	130**	≤ 50	*	Hemofilia
4	130**	≤ 50	*	Virose
5	122**	> 50	*	Hemofilia
6	122**	> 50	*	Câncer
7	122**	> 50	*	Virose
8	122**	> 50	*	Virose
9	130**	≤ 50	*	Câncer
10	130**	≤ 50	*	Câncer
11	130**	≤ 50	*	Hemofilia
12	130**	≤ 50	*	Virose

Limitações de ℓ-Diversidade

- Pode ser muito rigorosa

Limitações de ℓ-Diversidade

- Pode ser muito rigorosa
  - A sensibilidade dos atributos pode variar

Limitações de ℓ-Diversidade

- Pode ser muito rigorosa
  - A sensibilidade dos atributos pode variar
- Pode não ser suficiente

Limitações de ℓ-Diversidade

- Pode ser muito rigorosa
  - A sensibilidade dos atributos pode variar
- Pode não ser suficiente
  - Permite a inferência de atributos sensíveis, e em alguns casos com elevada probabilidade

Limitações de ℓ-Diversidade

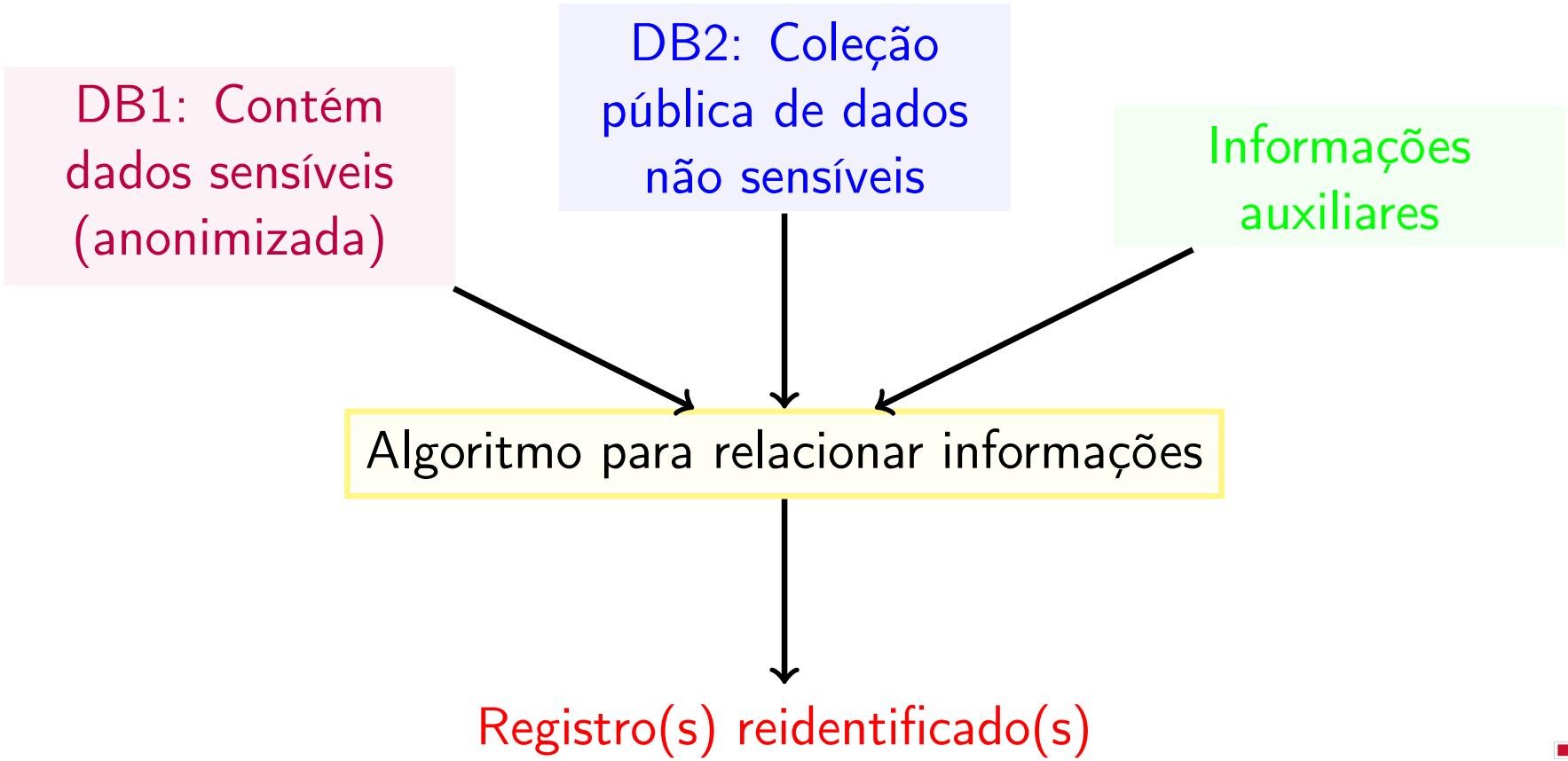
- Pode ser muito rigorosa
  - A sensibilidade dos atributos pode variar
- Pode não ser suficiente
  - Permite a inferência de atributos sensíveis, e em alguns casos com elevada probabilidade

t-Proximidade (Li et al., 2007)

Definição

Uma base de dados é dita t-Próxima se cada agrupamento de registros apresentar uma distribuição de atributos sensíveis próxima, a no máximo uma distância t, da distribuição geral

Composicionalidade



Composicionalidade

	Não Sensível			Sensível
	CEP	Idade	País	Condição
1	130**	< 30	*	AIDS
2	130**	< 30	*	Cardíaco
3	130**	< 30	*	Virose
4	130**	< 30	*	Virose
5	130**	≥ 40	*	Câncer
6	130**	≥ 40	*	Cardíaco
7	130**	≥ 40	*	Virose
8	130**	≥ 40	*	Virose
9	130**	3*	*	Câncer
10	130**	3*	*	Câncer
11	130**	3*	*	Câncer
12	130**	3*	*	Câncer

	Não Sensível			Sensível
	CEP	Idade	País	Condição
1	130**	< 35	*	AIDS
2	130**	< 35	*	Tuberculose
3	130**	< 35	*	Gripe
4	130**	< 35	*	Tuberculose
5	130**	< 35	*	Câncer
6	130**	< 35	*	Câncer
7	130**	≥ 35	*	Câncer
8	130**	≥ 35	*	Câncer
9	130**	≥ 35	*	Câncer
10	130**	≥ 35	*	Tuberculose
11	130**	≥ 35	*	Virose
12	130**	≥ 35	*	Virose

Alice tem 28 anos de idade, mora no CEP 13012, e visita ambos os hospitais

Netflix Prize Dataset (2008)

- 500.000 registros anônimos de classificações de filmes

Netflix Prize Dataset (2008)

- 500.000 registros anônimos de classificações de filmes
- Objetivo era fomentar pesquisas científicas

Netflix Prize Dataset (2008)

- 500.000 registros anônimos de classificações de filmes
- Objetivo era fomentar pesquisas científicas
- Narayanan & Shmatikov cruzaram as informações com perfis públicos no IMBD (*The Internet Movie Database*)

Netflix Prize Dataset (2008)

- 500.000 registros anônimos de classificações de filmes
- Objetivo era fomentar pesquisas científicas
- Narayanan & Shmatikov cruzaram as informações com perfis públicos no IMBD (*The Internet Movie Database*)
- Saber apenas algumas preferências (2-8 filmes, imprecisas) de um assinante foram suficientes para realizar a reidentificação

Netflix Prize Dataset (2008)

- 500.000 registros anônimos de classificações de filmes
- Objetivo era fomentar pesquisas científicas
- Narayanan & Shmatikov cruzaram as informações com perfis públicos no IMBD (*The Internet Movie Database*)
- Saber apenas algumas preferências (2-8 filmes, imprecisas) de um assinante foram suficientes para realizar a reidentificação
- Foi possível inferir posicionamento político e outras informações sensíveis

Anonimização

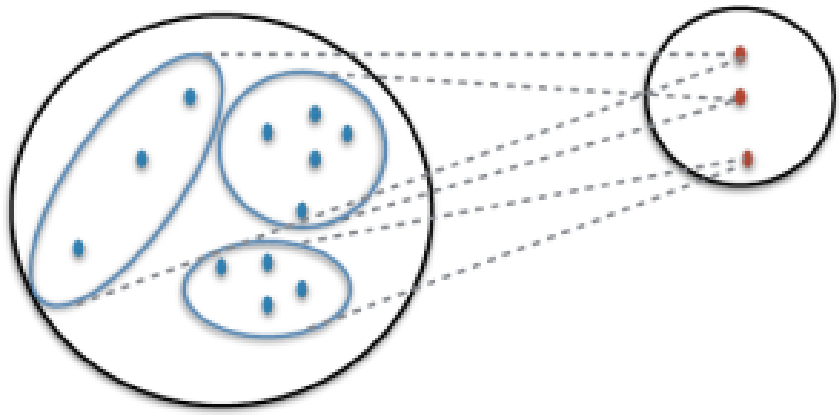
Métodos Probabilísticos

Ataques Composicionais

- Problema de métodos determinísticos

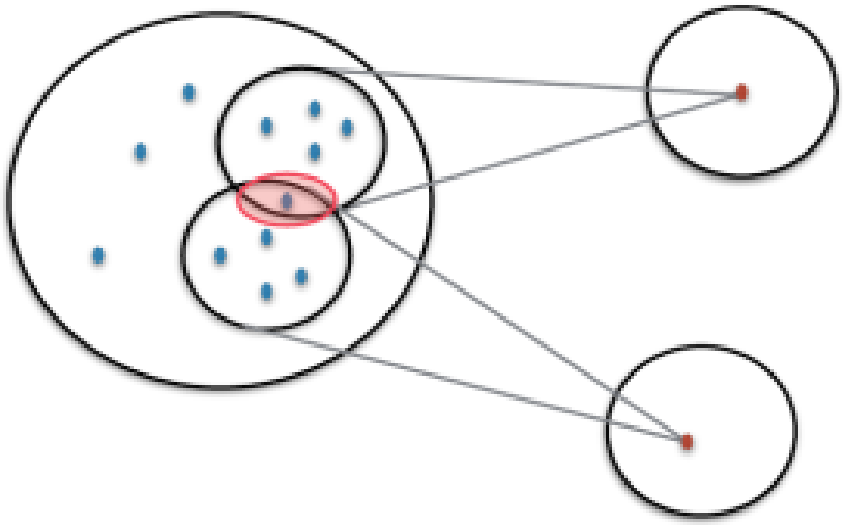
Ataques Composicionais

- Problema de métodos determinísticos
- Cada observável corresponde a um conjunto de segredos



Ataques Composicionais

- Problema de métodos determinísticos
- Cada observável corresponde a um conjunto de segredos
- Combinar observações diferentes permite determinar intersecções no domínio dos segredos

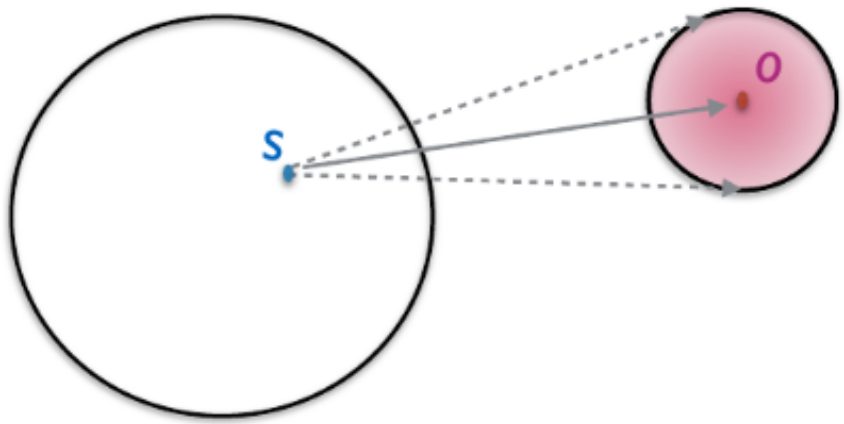


Visão Geral

- Todo segredo pode gerar qualquer observável de acordo com uma distribuição de probabilidade

Visão Geral

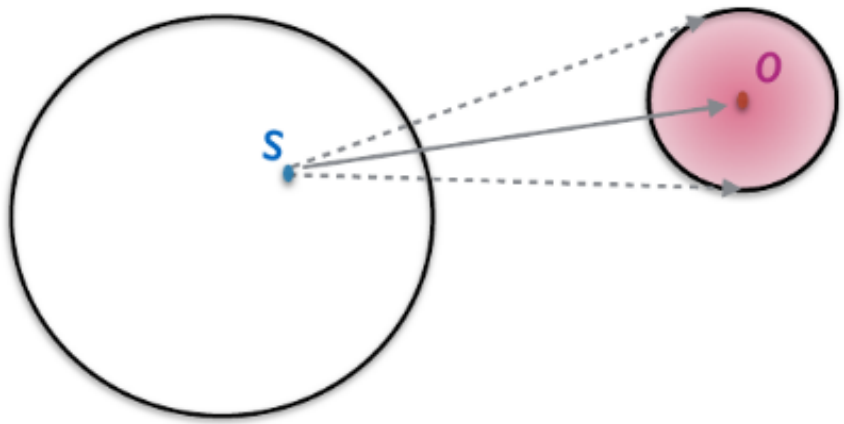
- Todo segredo pode gerar qualquer observável de acordo com uma distribuição de probabilidade



Visão Geral

- Todo segredo pode gerar qualquer observável de acordo com uma distribuição de probabilidade
- Teorema de Bayes:

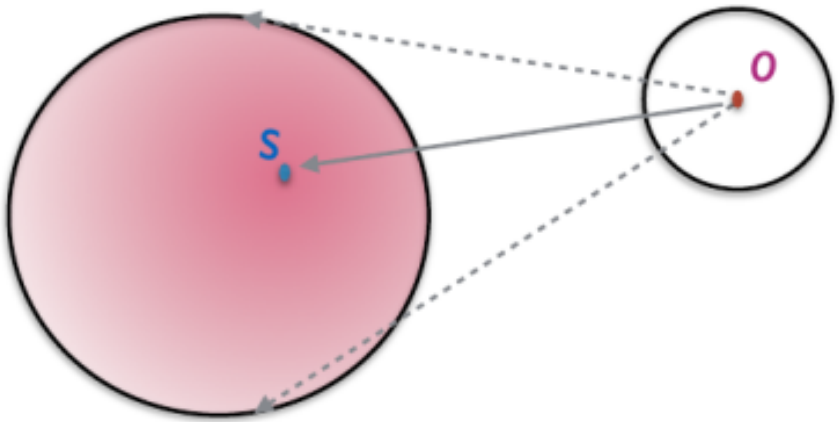
$$p(s|o) = \frac{p(s)}{p(o)} \cdot p(o|s)$$



Visão Geral

- Todo segredo pode gerar qualquer observável de acordo com uma distribuição de probabilidade
- Teorema de Bayes:

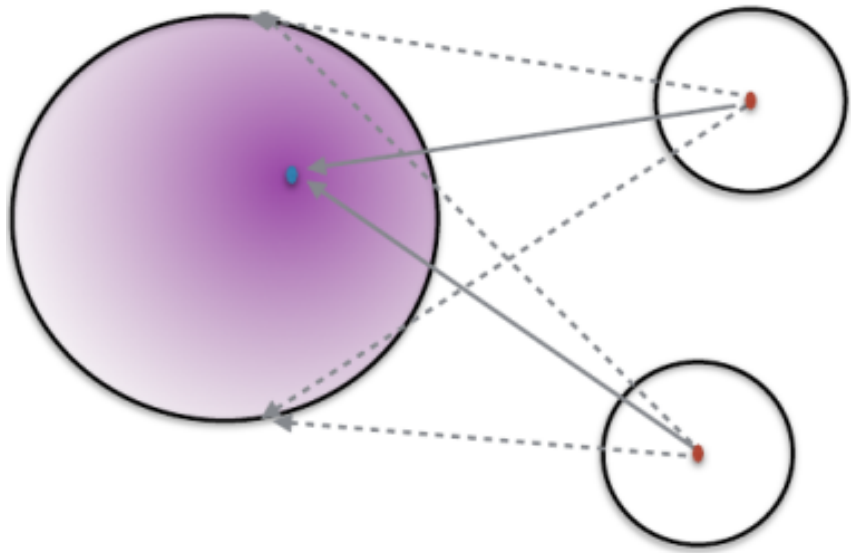
$$p(s|o) = \frac{p(s)}{p(o)} \cdot p(o|s)$$



Visão Geral

- Todo segredo pode gerar qualquer observável de acordo com uma distribuição de probabilidade
- Teorema de Bayes:

$$p(s|o) = \frac{p(s)}{p(o)} \cdot p(o|s)$$



Exemplo

Nome	Idade	Doença
Alice	30	não
Bob	30	não
Carl	40	não
Don	40	sim
Ellie	50	não
Frank	50	sim

Idade mínima com doença:

Exemplo

Nome	Idade	Doença
Alice	30	não
Bob	30	não
Carl	40	não
Don	40	sim
Ellie	50	não
Frank	50	sim

Idade mínima com doença:

- 30 com probabilidade 1/4

Exemplo

Nome	Idade	Doença
Alice	30	não
Bob	30	não
Carl	40	não
Don	40	sim
Ellie	50	não
Frank	50	sim

Idade mínima com doença:

- 30 com probabilidade 1/4
- 40 com probabilidade 1/2

Exemplo

Nome	Idade	Doença
Alice	30	não
Bob	30	não
Carl	40	não
Don	40	sim
Ellie	50	não
Frank	50	sim

Idade mínima com doença:

- 30 com probabilidade 1/4
- 40 com probabilidade 1/2
- 50 com probabilidade 1/4

Exemplo

Nome	Idade	Doença
Alice	30	não
Bob	30	não
Carl	40	não
Don	40	sim
Ellie	50	não
Frank	50	sim

Idade mínima com doença:

- 30 com probabilidade 1/4
- 40 com probabilidade 1/2
- 50 com probabilidade 1/4

Alice	Bob
Carl	Don
Ellie	Frank

Exemplo

Nome	Peso	Doença
Alice	60	não
Bob	90	não
Carl	90	não
Don	100	sim
Ellie	60	não
Frank	100	sim

Peso mínimo com doença:

Exemplo

Nome	Peso	Doença
Alice	60	não
Bob	90	não
Carl	90	não
Don	100	sim
Ellie	60	não
Frank	100	sim

Peso mínimo com doença:

- 60 com probabilidade 1/7

Exemplo

Nome	Peso	Doença
Alice	60	não
Bob	90	não
Carl	90	não
Don	100	sim
Ellie	60	não
Frank	100	sim

Peso mínimo com doença:

- 60 com probabilidade 1/7
- 90 com probabilidade 2/7

Exemplo

Nome	Peso	Doença
Alice	60	não
Bob	90	não
Carl	90	não
Don	100	sim
Ellie	60	não
Frank	100	sim

Peso mínimo com doença:

- 60 com probabilidade 1/7
- 90 com probabilidade 2/7
- 100 com probabilidade 4/7

Exemplo

Nome	Peso	Doença
Alice	60	não
Bob	90	não
Carl	90	não
Don	100	sim
Ellie	60	não
Frank	100	sim

Peso mínimo com doença:

- 60 com probabilidade 1/7
- 90 com probabilidade 2/7
- 100 com probabilidade 4/7

Alice	Bob
Carl	Don
Ellie	Frank

Exemplo

Nome	Idade	Doença
Alice	30	não
Bob	30	não
Carl	40	não
Don	40	sim
Ellie	50	não
Frank	50	sim

Nome	Peso	Doença
Alice	60	não
Bob	90	não
Carl	90	não
Don	100	sim
Ellie	60	não
Frank	100	sim

Alice	Bob
Carl	Don
Ellie	Frank

Observações

- Deve-se escolher a distribuição de probabilidade com cuidado

Observações

- Deve-se escolher a distribuição de probabilidade com cuidado
- O mecanismo deve proporcionar um equilíbrio entre **privacidade** e **utilidade**

Definições

Distância de Hamming

A Distância de Hammning entre dois bancos de dados x_1 e x_2 é igual ao número de registros que diferem entre x_1 e x_2

Definições

Distância de Hamming

A Distância de Hammning entre dois bancos de dados x_1 e x_2 é igual ao número de registros que diferem entre x_1 e x_2

Bancos de Dados Adjacentes

Dois bancos de dados x_1 e x_2 são adjacentes se a Distância de Hammning entre eles é igual a um. Denotamos essa propriedade como $x_1 \sim x_2$

Definições

Distância de Hamming

A Distância de Hammning entre dois bancos de dados x_1 e x_2 é igual ao número de registros que diferem entre x_1 e x_2

Bancos de Dados Adjacentes

Dois bancos de dados x_1 e x_2 são adjacentes se a Distância de Hammning entre eles é igual a um. Denotamos essa propriedade como $x_1 \sim x_2$

- x_1 e x_2 diferem em relação a apenas um indivíduo, *i.e.*

Definições

Distância de Hamming

A Distância de Hammning entre dois bancos de dados x_1 e x_2 é igual ao número de registros que diferem entre x_1 e x_2

Bancos de Dados Adjacentes

Dois bancos de dados x_1 e x_2 são adjacentes se a Distância de Hammning entre eles é igual a um. Denotamos essa propriedade como $x_1 \sim x_2$

- x_1 e x_2 diferem em relação a apenas um indivíduo, *i.e.*
  - ou o indivíduo foi adicionado a apenas uma das bases

Definições

Distância de Hamming

A Distância de Hammning entre dois bancos de dados x_1 e x_2 é igual ao número de registros que diferem entre x_1 e x_2

Bancos de Dados Adjacentes

Dois bancos de dados x_1 e x_2 são adjacentes se a Distância de Hammning entre eles é igual a um. Denotamos essa propriedade como $x_1 \sim x_2$

- x_1 e x_2 diferem em relação a apenas um indivíduo, *i.e.*
  - ou o indivíduo foi adicionado a apenas uma das bases
  - ou removido de apenas uma das bases

Definições

Distância de Hamming

A Distância de Hammning entre dois bancos de dados x_1 e x_2 é igual ao número de registros que diferem entre x_1 e x_2

Bancos de Dados Adjacentes

Dois bancos de dados x_1 e x_2 são adjacentes se a Distância de Hammning entre eles é igual a um. Denotamos essa propriedade como $x_1 \sim x_2$

- x_1 e x_2 diferem em relação a apenas um indivíduo, *i.e.*
  - ou o indivíduo foi adicionado a apenas uma das bases
  - ou removido de apenas uma das bases
  - ou teve seu valor alterado em apenas uma das bases

Privacidade ϵ-Diferencial (Dwork, 2006)

Definição

Uma base de dados é dita Privada ϵ-Diferencial se para todos os bancos de dados x, x' adjacentes, e para todo $z \in Z$, a resposta reportada, temos que:

$$\frac{p(K = z|X = x)}{p(K = z|X = x')} \leq e^{\epsilon}$$

Privacidade ϵ-Diferencial (Dwork, 2006)

Definição

Uma base de dados é dita Privada ϵ-Diferencial se para todos os bancos de dados x, x' adjacentes, e para todo $z \in Z$, a resposta reportada, temos que:

$$\frac{p(K = z|X = x)}{p(K = z|X = x')} \leq e^{\epsilon}$$

Interpretação

A presença ou não de informações sobre um indivíduo na base de dados não altera significativamente os resultados obtidos

Propriedades da Privacidade Diferencial

- Independência da *prior*, a distribuição de probabilidade sobre os segredos antes de se consultar o banco de dados

Propriedades da Privacidade Diferencial

- Independência da *prior*, a distribuição de probabilidade sobre os segredos antes de se consultar o banco de dados
  - Independe do adversário

Propriedades da Privacidade Diferencial

- Independência da *prior*, a distribuição de probabilidade sobre os segredos antes de se consultar o banco de dados
  - Independe do adversário
- Robustez quanto à Composicionalidade

Propriedades da Privacidade Diferencial

- Independência da *prior*, a distribuição de probabilidade sobre os segredos antes de se consultar o banco de dados
  - Independe do adversário
- Robustez quanto à Composicionalidade
  - Dados dois mecanismos Privados ϵ_1-Diferencial e ϵ_2-Diferencial, a composição é um mecanismo $(\epsilon_1 + \epsilon_2)$-Diferencial

Propriedades da Privacidade Diferencial

- Independência da *prior*, a distribuição de probabilidade sobre os segredos antes de se consultar o banco de dados
  - Independe do adversário
- Robustez quanto à Composicionalidade
  - Dados dois mecanismos Privados ϵ_1-Diferencial e ϵ_2-Diferencial, a composição é um mecanismo $(\epsilon_1 + \epsilon_2)$-Diferencial
  - A privacidade diminui linearmente com o número de consultas

Propriedades da Privacidade Diferencial

- Independência da *prior*, a distribuição de probabilidade sobre os segredos antes de se consultar o banco de dados
  - Independe do adversário
- Robustez quanto à Composicionalidade
  - Dados dois mecanismos Privados ϵ_1-Diferencial e ϵ_2-Diferencial, a composição é um mecanismo $(\epsilon_1 + \epsilon_2)$-Diferencial
  - A privacidade diminui linearmente com o número de consultas
  - É possível definir um **orçamento de privacidade**, ou seja, um máximo aceitável de violação da privacidade

Exemplos de Privacidade Diferencial

Qual a altura média?
A distribuição varia entre
50 cm e 250 cm

	Altura (cm)
Alice	140
Bob	180
Carol	160
Daniel	120 (p) 190 ($1-p$)

Exemplos de Privacidade Diferencial

Qual a altura média?
A distribuição varia entre 50 cm e 250 cm

	Altura (cm)
Alice	140
Bob	180
Carol	160
Daniel	120 (p) 190 ($1-p$)

Um mecanismo que **sempre** reporta a resposta verdadeira (150 cm) não é privado, qualquer que seja o ϵ

$$\frac{p(\textit{média} = 150 | \textit{Daniel} = 120)}{p(\textit{média} = 150 | \textit{Daniel} = 190)} = \frac{1}{0} = e^{\infty}$$

Exemplos de Privacidade Diferencial

Qual a altura média?
A distribuição varia entre 50 cm e 250 cm

	Altura (cm)
Alice	140
Bob	180
Carol	160
Daniel	120 (p) 190 ($1-p$)

Um mecanismo que **sempre** reporta uma resposta errada (168 cm) é completamente privado ($\epsilon = 0$), mas completamente inútil

$$\frac{p(\mathit{média} = 168 | \mathit{Daniel} = x)}{p(\mathit{média} = 168 | \mathit{Daniel} = x')} = \frac{1}{1} = e^0$$

$$\frac{p(\mathit{média} = z | \mathit{Daniel} = x)}{p(\mathit{média} = z | \mathit{Daniel} = x')} = \frac{0}{0} \quad (z \neq 168)$$

Exemplos de Privacidade Diferencial

Qual a altura média?
A distribuição varia entre 50 cm e 250 cm

	Altura (cm)
Alice	140
Bob	180
Carol	160
Daniel	120 (p) 190 ($1-p$)

Um mecanismo que **sempre** reporta 100 cm se a resposta verdadeira (150 cm) for menor ou igual a 150 cm, ou reporta 200 cm caso contrário, não é privado, qualquer que seja o ϵ

$$\frac{p(\mathit{média} = 100 | \mathit{Daniel} = 120)}{p(\mathit{média} = 100 | \mathit{Daniel} = 190)} = \frac{1}{0} = e^{\infty}$$

Exemplos de Privacidade Diferencial

Qual a altura média?
A distribuição varia entre 50 cm e 250 cm

	Altura (cm)
Alice	140
Bob	180
Carol	160
Daniel	120 (p) 190 ($1-p$)

O mecanismo que reporta a resposta verdadeira com probabilidade $\epsilon/(200+\epsilon)$ e todos os outros inteiros no intervalo $[50, 250]$ com probabilidade $1/(200+\epsilon)$, é privado ϵ-diferencial

Estado da Arte

Google, Microsoft, e Apple

Privacidade Diferencial Local

Definição

A randomização e a adição de ruído são realizadas pelo software em execução no dispositivo do usuário

Privacidade Diferencial Local

Exemplos

As seguintes empresas privadas reportam utilizar Privacidade Diferencial Local para coleta de Telemetria

Privacidade Diferencial Local

Exemplos

As seguintes empresas privadas reportam utilizar Privacidade Diferencial Local para coleta de Telemetria

- Google (Chrome): RAPPOR (Randomized Aggregatable Privacy-Preserving Ordinal Response)

Privacidade Diferencial Local

Exemplos

As seguintes empresas privadas reportam utilizar Privacidade Diferencial Local para coleta de Telemetria

- Google (Chrome): RAPPOR (Randomized Aggregatable Privacy-Preserving Ordinal Response)
- Apple: iOS 10 em diante

Privacidade Diferencial Local

Exemplos

As seguintes empresas privadas reportam utilizar Privacidade Diferencial Local para coleta de Telemetria

- Google (Chrome): RAPPOR (Randomized Aggregatable Privacy-Preserving Ordinal Response)
- Apple: iOS 10 em diante
- Microsoft: Windows 10

Privacidade Diferencial Local

Exemplos

As seguintes empresas privadas reportam utilizar Privacidade Diferencial Local para coleta de Telemetria

- Google (Chrome): RAPPOR (Randomized Aggregatable Privacy-Preserving Ordinal Response)
- Apple: iOS 10 em diante
- Microsoft: Windows 10

O objetivo final é aumentar a aceitação pública da coleta de dados

Privacidade Diferencial Local

Exemplos

As seguintes empresas privadas reportam utilizar Privacidade Diferencial Local para coleta de Telemetria

- Google (Chrome): RAPPOR (Randomized Aggregatable Privacy-Preserving Ordinal Response)
- Apple: iOS 10 em diante
- Microsoft: Windows 10

O objetivo final é aumentar a aceitação pública da coleta de dados

Entretanto, para garantir a utilidade dos dados coletados, utilizam um orçamento de privacidade (ϵ) muito grande

Estado da Arte

US Census Bureau

Visão Geral

- Responsável pela realização dos Censos decenais

Visão Geral

- Responsável pela realização dos Censos decenais
- Os resultados são publicados agregados e servem de base para

Visão Geral

- Responsável pela realização dos Censos decenais
- Os resultados são publicados agregados e servem de base para
  - a repartição dos assentos na Câmara dos Deputados entre os estados

Visão Geral

- Responsável pela realização dos Censos decenais
- Os resultados são publicados agregados e servem de base para
  - a repartição dos assentos na Câmara dos Deputados entre os estados
  - a distribuição de mais de US$675 bilhões em fundos federais para estados e organizações locais

Visão Geral

- Responsável pela realização dos Censos decenais
- Os resultados são publicados agregados e servem de base para
  - a repartição dos assentos na Câmara dos Deputados entre os estados
  - a distribuição de mais de US$675 bilhões em fundos federais para estados e organizações locais
- Por Lei, a Confidencialidade deve ser garantida

Visão Geral

- Até o Censo de 1970, apenas desidentificava os registros

Visão Geral

- Até o Censo de 1970, apenas desidentificava os registros
- Entre os Censos de 1980 e de 2010, aplicou também métodos determinísticos

Visão Geral

- Até o Censo de 1970, apenas desidentificava os registros
- Entre os Censos de 1980 e de 2010, aplicou também métodos determinísticos
- Teorema da Reconstrução da Base de Dados (Dinur & Nissim, 2003)

Visão Geral

- Até o Censo de 1970, apenas desidentificava os registros
- Entre os Censos de 1980 e de 2010, aplicou também métodos determinísticos
- Teorema da Reconstrução da Base de Dados (Dinur & Nissim, 2003)
  - "Muitas estatísticas publicadas com muita precisão a partir de uma base de dados confidencial expõem a base de dados inteira com quase certeza"

Visão Geral

- Até o Censo de 1970, apenas desidentificava os registros
- Entre os Censos de 1980 e de 2010, aplicou também métodos determinísticos
- Teorema da Reconstrução da Base de Dados (Dinur & Nissim, 2003)
  - "Muitas estatísticas publicadas com muita precisão a partir de uma base de dados confidencial expõem a base de dados inteira com quase certeza"
  - Para se ter privacidade, é necessário adicionar uma perturbação de magnitude ao menos $\sqrt{N}$, onde N é o tamanho da população

2020 Census

- Métodos determinísticos foram descartados devido à legislação

2020 Census

- Métodos determinísticos foram descartados devido à legislação
- Entretanto, Privacidade Diferencial Local adiciona muito ruído, caso seja utilizada para garantir de fato a privacidade

2020 Census

- Métodos determinísticos foram descartados devido à legislação
- Entretanto, Privacidade Diferencial Local adiciona muito ruído, caso seja utilizada para garantir de fato a privacidade
- Todo o banco de dados deve ser processado de uma só vez para garantir melhor precisão

2020 Census

- Métodos determinísticos foram descartados devido à legislação
- Entretanto, Privacidade Diferencial Local adiciona muito ruído, caso seja utilizada para garantir de fato a privacidade
- Todo o banco de dados deve ser processado de uma só vez para garantir melhor precisão
- Todos os usos dos dados privados devem ser previamente considerados no orçamento de privacidade, ou seja, antes da publicação dos dados

2020 Census

- Métodos determinísticos foram descartados devido à legislação
- Entretanto, Privacidade Diferencial Local adiciona muito ruído, caso seja utilizada para garantir de fato a privacidade
- Todo o banco de dados deve ser processado de uma só vez para garantir melhor precisão
- Todos os usos dos dados privados devem ser previamente considerados no orçamento de privacidade, ou seja, antes da publicação dos dados
- Foi desenvolvido um novo método, *Top-Down*, que permite a criação das tabelas com informações agregadas sem grandes perdas de utilidade

2020 Census

- Métodos determinísticos foram descartados devido à legislação
- Entretanto, Privacidade Diferencial Local adiciona muito ruído, caso seja utilizada para garantir de fato a privacidade
- Todo o banco de dados deve ser processado de uma só vez para garantir melhor precisão
- Todos os usos dos dados privados devem ser previamente considerados no orçamento de privacidade, ou seja, antes da publicação dos dados
- Foi desenvolvido um novo método, *Top-Down*, que permite a criação das tabelas com informações agregadas sem grandes perdas de utilidade
- Novos resultados referentes ao Censo de 2010 foram publicados para servirem de exemplo da aplicação do novo método

2020 Census

- Métodos determinísticos foram descartados devido à legislação
- Entretanto, Privacidade Diferencial Local adiciona muito ruído, caso seja utilizada para garantir de fato a privacidade
- Todo o banco de dados deve ser processado de uma só vez para garantir melhor precisão
- Todos os usos dos dados privados devem ser previamente considerados no orçamento de privacidade, ou seja, antes da publicação dos dados
- Foi desenvolvido um novo método, *Top-Down*, que permite a criação das tabelas com informações agregadas sem grandes perdas de utilidade
- Novos resultados referentes ao Censo de 2010 foram publicados para servirem de exemplo da aplicação do novo método
- US National Census Day: 1º de Abril de 2020

Resumo

Resumo

Governos e empresas privadas coletam e continuarão coletando dados pessoais

INSCRYPT
Laboratory of Information Security, Cryptography, Privacy, and Transparency

Governos e empresas privadas coletam e continuarão coletando dados pessoais

- É importante que essas entidades garantam a segurança dessas informações

INSCRYPT
Laboratory of Information Security, Cryptography, Privacy, and Transparency

Resumo

Governos e empresas privadas coletam e continuarão coletando dados pessoais

- É importante que essas entidades garantam a segurança dessas informações
  - Controle de acesso e Criptografia são muito importantes, mas não resolvem todos os problemas

INSCRYPT
Laboratory of Information Security, Cryptography, Privacy, and Transparency

Resumo

Governos e empresas privadas coletam e continuarão coletando dados pessoais

- É importante que essas entidades garantam a segurança dessas informações
  - Controle de acesso e Criptografia são muito importantes, mas não resolvem todos os problemas
- É importante que essas entidades continuem divulgando dados, mas de forma responsável

INSCRYPT
Laboratory of Information Security, Cryptography, Privacy, and Transparency

Resumo

Governos e empresas privadas coletam e continuarão coletando dados pessoais

- É importante que essas entidades garantam a segurança dessas informações
  - Controle de acesso e Criptografia são muito importantes, mas não resolvem todos os problemas
- É importante que essas entidades continuem divulgando dados, mas de forma responsável
  - Muitos métodos de Anonimização já foram propostos, mas diversos problemas permanecem em aberto

INSCRYPT
Laboratory of Information Security, Cryptography, Privacy, and Transparency

Resumo

Governos e empresas privadas coletam e continuarão coletando dados pessoais

- É importante que essas entidades garantam a segurança dessas informações
  - Controle de acesso e Criptografia são muito importantes, mas não resolvem todos os problemas
- É importante que essas entidades continuem divulgando dados, mas de forma responsável
  - Muitos métodos de Anonimização já foram propostos, mas diversos problemas permanecem em aberto
    - Como definir o conjunto de quasi-identificadores?

INSCRYPT
Laboratory of Information Security, Cryptography, Privacy, and Transparency

Resumo

Governos e empresas privadas coletam e continuarão coletando dados pessoais

- É importante que essas entidades garantam a segurança dessas informações
  - Controle de acesso e Criptografia são muito importantes, mas não resolvem todos os problemas
- É importante que essas entidades continuem divulgando dados, mas de forma responsável
  - Muitos métodos de Anonimização já foram propostos, mas diversos problemas permanecem em aberto
    - Como definir o conjunto de quasi-identificadores?
    - Como garantir um nível aceitável de utilidade?

INSCRYPT
Laboratory of Information Security, Cryptography, Privacy, and Transparency

Governos e empresas privadas coletam e continuarão coletando dados pessoais

- É importante que essas entidades garantam a segurança dessas informações
  - Controle de acesso e Criptografia são muito importantes, mas não resolvem todos os problemas
- É importante que essas entidades continuem divulgando dados, mas de forma responsável
  - Muitos métodos de Anonimização já foram propostos, mas diversos problemas permanecem em aberto
    - Como definir o conjunto de quasi-identificadores?
    - Como garantir um nível aceitável de utilidade?
    - Como anonimizar dados longitudinais no tempo?

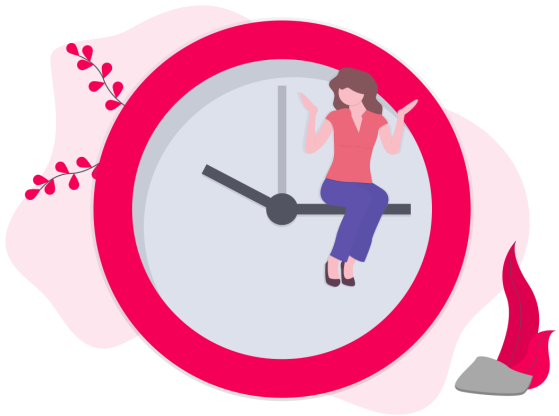
INSCRYPT
Laboratory of Information Security, Cryptography, Privacy, and Transparency

Como equilibrar Transparência e Privacidade?

INSCRYPT
Laboratory of Information Security, Cryptography, Privacy, and Transparency

Como equilibrar Transparência e Privacidade?

Lembrando que a Lei Geral de Proteção de Dados
entra em vigor em Agosto de 2020 no Brasil!



INSCRYPT
Laboratory of Information Security,
Cryptography, Privacy, and Transparency

Departamento de Ciência da Computação

Equipe da EVCOMP 2020

Prof. Mário Alvim e Prof. Annabelle McIver

Prof. Catuscia Palamidessi e Prof. Kostas Chatzikokolakis

Obrigado pela atenção!

ghn@nunesgh.com
nunesgh.com/sobre

INSCRYPT
Laboratory of Information Security, Cryptography, Privacy, and Transparency