463.5 De-Identification

Computer Security II
CS463/ECE424
University of Illinois

Security News of the Day

For The U.S. Census, Keeping Your Data Anonymous And Useful Is A Tricky Balance

Updated August 2, 2021 · 12:17 PM ET

https://www.npr.org/2021/05/19/993247101/for-the-u-s-census-keeping-your-data-anonymous-and-useful-is-a-tricky-balance

For decades, the bureau has stripped away names and addresses from census records before turning them into anonymized data. That information is broken down by race, ethnicity, age and sex to levels as detailed as a neighborhood.

statistics from the census. Advances in computing and access to voter registration lists and commercial data sets that can be cross-referenced have made it easier to trace purportedly anonymized information back to an individual person.

For a way out of this conundrum, the bureau has been building a new privacy protection system based on a mathematical concept known as differential privacy.

Preliminary tests of the bureau's new privacy protections have left many data users worried that their ability to use 2020 census statistics could be severely limited, particularly data about small geographic areas and minority groups within communities that many governments rely on for planning.

Outline

De-Identification
Privacy metrics
Privacy in practice

De-Identification

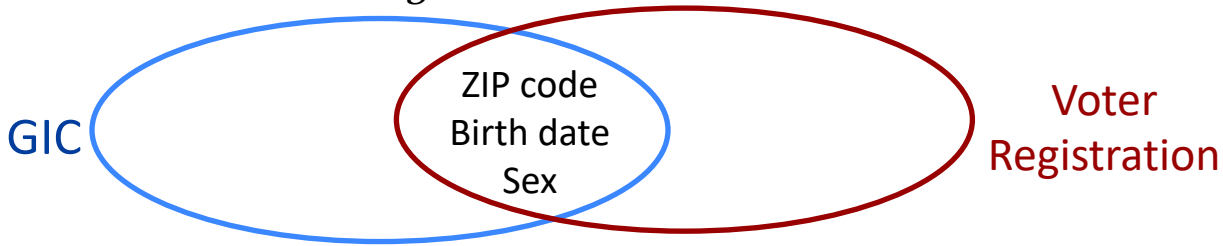
- Suppose we have a dataset we would like to release
- The dataset contains sensitive information about a set of individuals
- We want to protect the privacy of those individuals

Name	Sex	Age	Zip	Diagnosis (ICD-9)
Alice Smith	F	37	61821	037, 651
Bob Johnson	M	41	61820	823, 042
Carol Williams	F	24	61803	010, 650
Dan Jones	M	46	61706	823, 460
Elisabeth Brown	F	50	61824	945

- What about just removing the names?

GIC incident

- Group Insurance Commission (Massachusetts)
  - Release patient data of state employees (about 135,000 records)
  - De-identification of the dataset by removing names
- [Sweeney02] re-identification of the governor
  - linking the dataset with the voter registration list



- Uniqueness of demographics
  - (5-digit ZIP, birth date, sex) uniquely identifies over 87% of US population

[Sweeney02]: Sweeney, Latanya. "k-anonymity: A model for protecting privacy." International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10.05 (2002): 557-570.

AOL search logs incident

- AOL released search logs of 650,000 users in Aug 2006
  - De-identification of the dataset by using pseudonyms (a unique number for each customer)

- [New York Times 2006] Re-identified Thelma Arnold (user 4417749) through some of her searches:
  - "60 single men", "landscapers in Lilburn, Ga“
  - Also searched the names of some of her relatives, last name Arnold

- Class action lawsuit in Sept 2006
  - AOL’s CTO resigned, two employees were fired
  - Search logs can still be downloaded from mirrors


The New York Times

A Face Is Exposed for AOL Searcher No. 4417749

Netflix Prize incident

- Dataset containing movie ratings of 500,000 users
  - De-identification by removing identifiers, using a randomly assigned ID in place of the customer ID

- [NS08] Proposed new class of attacks target high dimensionality sparse datasets
  - Using 8 movie ratings (2 can be wrong) and dates (with up to 14 days error), 99% of users are uniquely identifiable
  - (Proof of concept) Re-identified 2 users by linking the Netflix dataset to IMDb using a sample of 50 IMDb users

[NS08]: Narayanan and Shmatikov. "Robust de-anonymization of large sparse datasets." IEEE S&P 2008.

Re-identification Vectors

- External Knowledge
  - E.g., voter registration, marriage registries
- Unredacted free text
  - Can contain arbitrary data
- High dimensionality, sparsity
  - More features
  - Large distance between data points
  - More likely to be unique

Types of disclosure

- Identification disclosure
  - Reveals the target individual's record
- Attribute disclosure
  - Reveals one or more (possibly sensitive) attributes about the target individual
  - Can occur even
    - without identification disclosure
    - if the target individual's record is not in the dataset (e.g., "smoking causes cancer")
- Membership disclosure
  - Reveals whether the target individual's record was included in the dataset

k-anonymity: Hiding in a Crowd of K People

- [Sweeney02] k-anonymity
  - Quasi-identifiers: attributes that can be used for linking with external information (e.g., ZIP code, sex, birth date)
  - To satisfy k-anonymity: any sequence of quasi-identifiers must appear in <u>at least</u> k records

Name	Sex	Age	Zip	Diagnosis (ICD-9)
Alice Smith	F	37	61821	037, 651
Bob Johnson	M	41	61820	823, 042
Carol Williams	F	24	61803	010, 650
Dan Jones	M	46	61706	823, 460
Elisabeth Brown	F	50	61824	945

Quasi-identifiers

Satisfying k-anonymity

- Generalization
  - E.g., ZIP codes – 61802 -> 61XXX
  - E.g., Age – 47 -> [40, 49]
- Suppression:
  - E.g., names
- Is this 2-anonymous?

	Sex	Age	Zip	Diagnosis
	*	[30-39]	61XXX	Broken Leg
	*	[40-49]	61XXX	Cancer
	*	[40-49]	61XXX	Cancer
	*	[30-39]	61XXX	Tuberculosis
No!	*	[20-29]	61XXX	Heart Condition

Satisfying k-anonymity

- Generalization
  - E.g., ZIP codes – 61802 -> 61XXX
  - E.g., Age – 47 -> [40, 49]
- Suppression:
  - E.g., names
- Is this 2-anonymous?

How about now? → YES!

Sex	Age	Zip	Diagnosis
*	[30-39]	61XXX	Broken Leg
*	[40-49]	61XXX	Cancer
*	[40-49]	61XXX	Cancer
*	[30-39]	61XXX	Tuberculosis

Other syntactic metrics

- k-anonymity does not prevent attribute disclosure
  - E.g., if there is a quasi-identifier group among which all records contain an attribute that has a single value

Other syntactic metrics

- k-anonymity does not prevent attribute disclosure
  - E.g., if there is a quasi-identifier group among which all records contain an attribute that has a single value
- l-diversity
  - Within each quasi-identifier group, there must be at least l distinct values for each attribute

Sex	Age	Zip	Diagnosis
*	[30-39]	61XXX	Cancer
*	[30-39]	61XXX	Cancer
*	[30-39]	61XXX	Cancer
*	[30-39]	61XXX	Broken Leg
*	[30-39]	61XXX	Cancer

5-anonymous
2-diverse

Other syntactic metrics

- k-anonymity does not prevent attribute disclosure
  - E.g., if there is a quasi-identifier group among which all records contain an attribute that has a single value
- l-diversity
  - Within each quasi-identifier group, there must be at least l distinct values for each attribute
- t-closeness
  - The distance between the distribution of attributes within a quasi-identifier group and the overall distribution should not exceed t

Example

- For what values of k and l is the dataset k-anonymous and l-diverse?

Sex	Age	Diagnosis
M	[40-49]	Cancer
F	[40-49]	HIV
M	[30-39]	Asthma
F	[30-39]	Influenza
F	[30-39]	Cancer
M	[30-39]	Broken Leg
F	[30-39]	Tuberculosis
M	[40-49]	Tuberculosis
F	[40-49]	HIV

Quasi-identifier group	records	sensitive values
(M, [30-39])	2	2
(M, [40-49])	2	2
(F, [30-39])	3	3
(F, [40-49])	2	1

Example

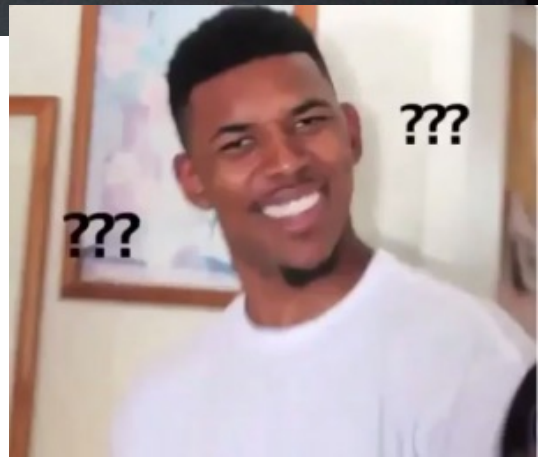
- For what values of k and l is the dataset k-anonymous and l-diverse?

Sex	Age	Diagnosis
M	[40-49]	Cancer
F	[40-49]	HIV
M	[30-39]	Asthma
F	[30-39]	Influenza
F	[30-39]	Cancer
M	[30-39]	Broken Leg
F	[30-39]	Tuberculosis
M	[40-49]	Tuberculosis
F	[40-49]	HIV

Quasi-identifier group	records	sensitive values
(M, [30-39])	2	2
(M, [40-49])	2	2
(F, [30-39])	3	3
(F, [40-49])	2	1

2-anonymous
1-diverse

OK, How about Privacy?



Differential Privacy [Dwork06]

- Intuition: what can be learned from accessing the database is (roughly) the same regardless of whether an individual is in the database.

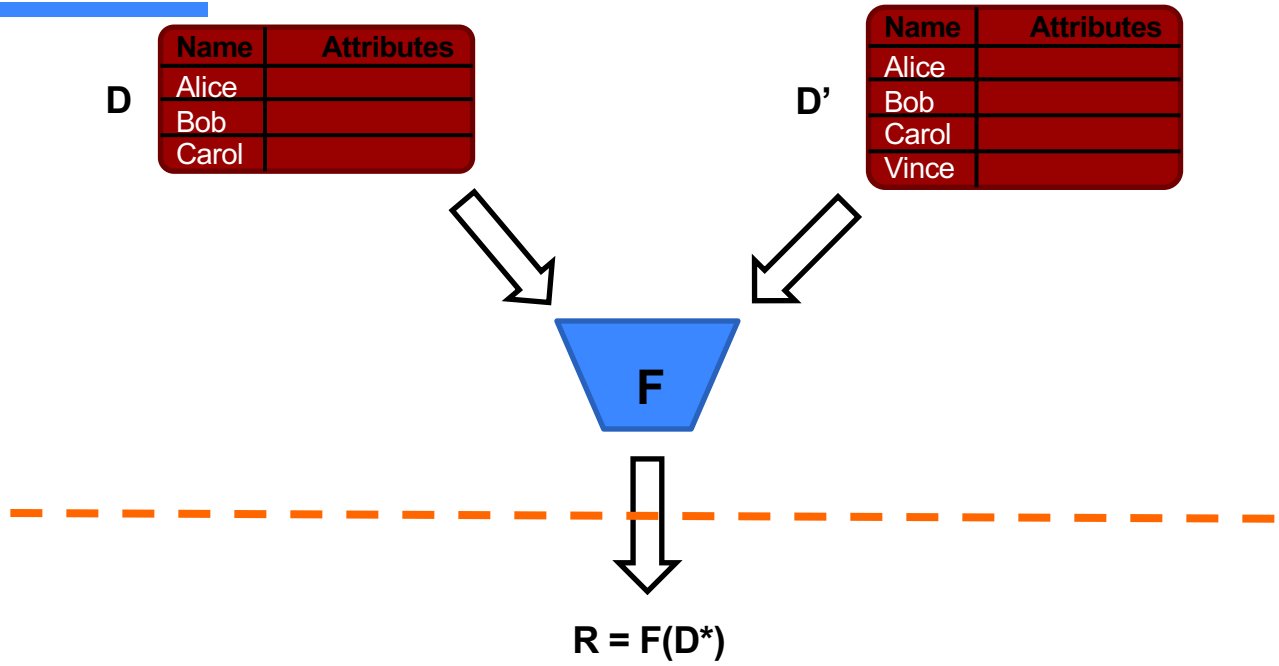
- For any two datasets D and D' differing in a single record. A computation F is ϵ-differentially private for some $\epsilon > 0$, if for all $S \subseteq \text{Range}(F)$, we have:

Privacy budget

$$\mathbb{P}[F(D) \in S] \leq e^{\epsilon} \cdot \mathbb{P}[F(D') \in S]$$

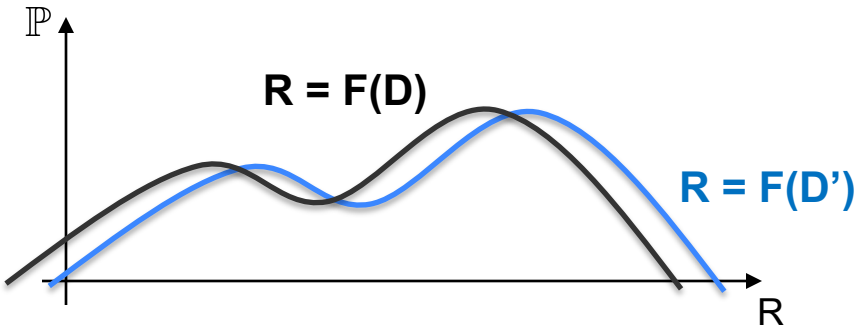
D' includes the user's record, D doesn't

One way to think about it



Probability distribution over R should be "roughly" the same whether D = D, or D* = D'.*

Slightly More Formally



$$e^{-\epsilon} \leq \frac{\mathbb{P}[R|D^*=D]}{\mathbb{P}[R|D^*=D']} \leq e^{\epsilon}$$

- The probability distribution is over the random coins of F.
- Note: $e^{\epsilon} \approx 1+\epsilon$, for a small $\epsilon > 0$

Privatization

- Idea: add noise
  - What noise distribution should we use?
  - How much noise to add?

- The key concept is sensitivity of f, the function we want to compute

- Generic way to get ϵ-differential privacy: Laplacian mechanism

Sensitivity

$$\Delta f = \max_{D,D'} |f(D) - f(D')|$$

- Sensitivity measures how much an individual record can change the output, i.e., f(D*), *in the worst case*

- E.g., count() function has sensitivity of 1
- E.g., average() may have a high sensitivity.

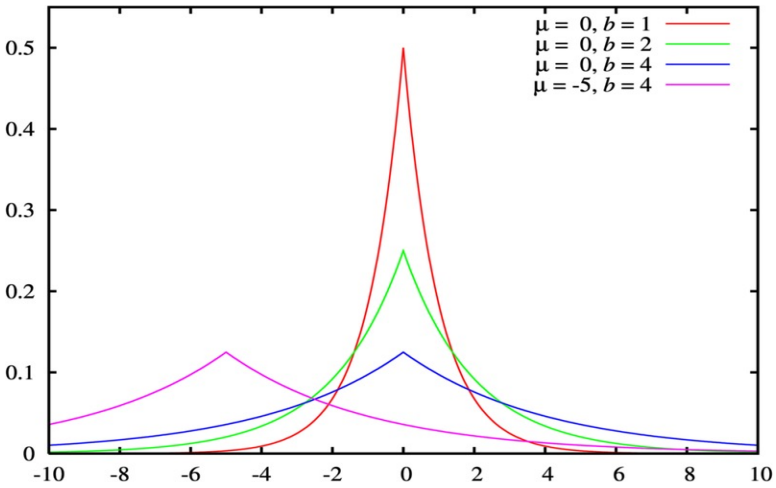
Laplacian Mechanism

- Add noise from Laplace distribution
- That is, release: $$f(D^*) + Lap(\frac{\Delta f}{\epsilon})$$

Laplace Distribution with mean 0

$$\mathbb{P}[x] = \frac{1}{2b} e^{-\frac{|x|}{b}}$$

The Laplace mechanism is $(\epsilon, 0)$-differentially private



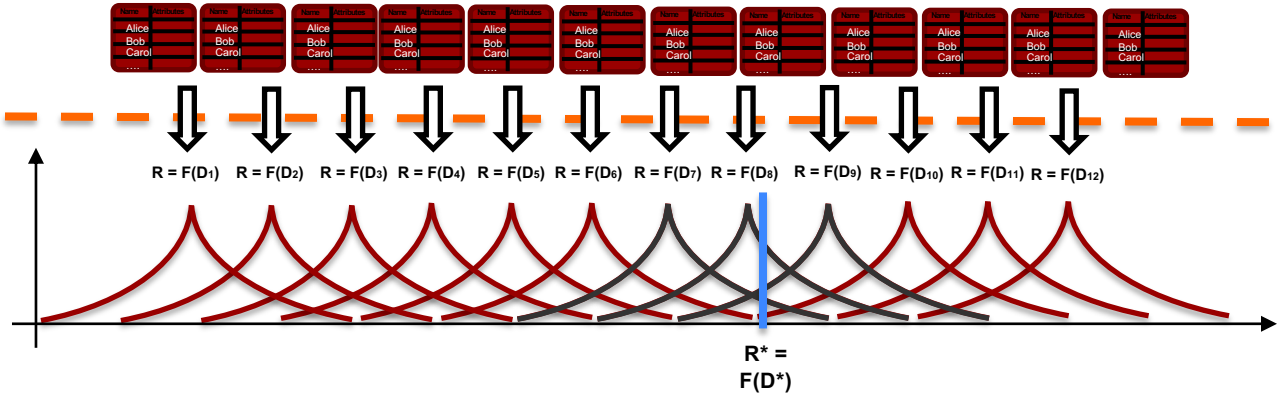
From Wikipedia.

Why does it work?

- Intuitively:

Why does it work?

- Intuitively:

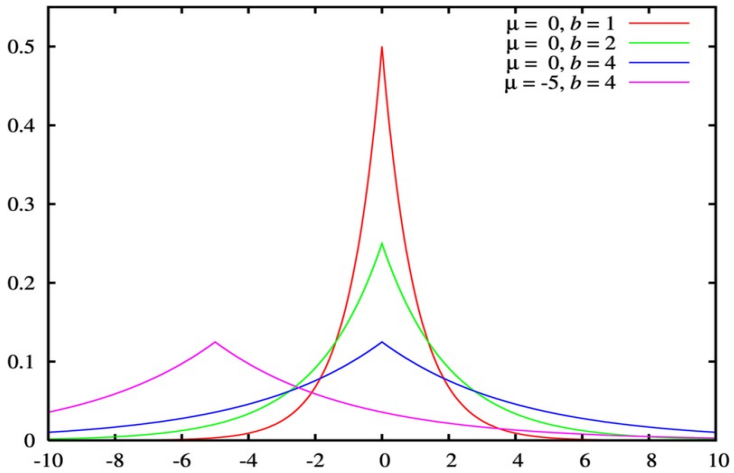


Why does it work?

- Mathematically:

Laplace Distribution with mean 0

$$\mathbb{P}[x] = \frac{1}{2b} e^{-\frac{|x|}{b}}$$



From Wikipedia.

$$\frac{\mathbb{P}[F(D) \in S]}{\mathbb{P}[F(D') \in S]} = \frac{\int_S \mathbb{P}[R = r|D]}{\int_S \mathbb{P}[R = r|D']} \leq \max_{r \in S} \frac{\mathbb{P}[R = r|D]}{\mathbb{P}[R = r|D']}$$

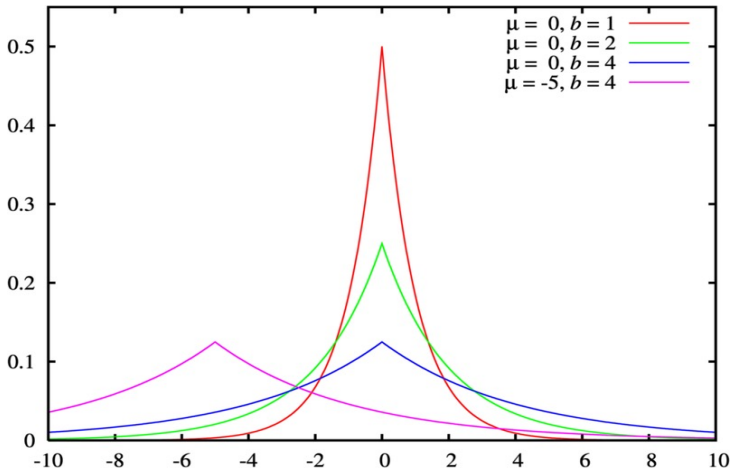
See Aaron Roth slides for more details: http://www.cis.upenn.edu/~aaroth/courses/slides/Lecture3.pdf

Why does it work?

- Mathematically:

Laplace Distribution with mean 0

$$\mathbb{P}[x] = \frac{1}{2b} e^{-\frac{|x|}{b}}$$



From Wikipedia.

$$\frac{\mathbb{P}[F(D) \in S]}{\mathbb{P}[F(D') \in S]} = \frac{\int_S \mathbb{P}[R = r|D]}{\int_S \mathbb{P}[R = r|D']} \leq \max_{r \in S} \frac{\mathbb{P}[R = r|D]}{\mathbb{P}[R = r|D']}$$

$$\frac{\mathbb{P}[R|D* = D]}{\mathbb{P}[R|D* = D']} = \frac{\cancel{\frac{\epsilon}{2\Delta f}} \exp\left(-\frac{\epsilon}{\Delta f}|R - f(D)|\right)}{\cancel{\frac{\epsilon}{2\Delta f}} \exp\left(-\frac{\epsilon}{\Delta f}|R - f(D')|\right)} \leq e^{\epsilon}$$

See Aaron Roth slides for more details: http://www.cis.upenn.edu/~aaroth/courses/slides/Lecture3.pdf

Composition

- What about multiple queries?
- Sequential composition theorem:
  - Making $t \geq 1$ ϵ-differentially private queries gives us $t\epsilon$-differential privacy

- In practice:
  1. Set a privacy budget ϵ
  2. Each query uses ϵ' of the remaining budget
  3. Once the privacy budget exceeded, stop answering

Using Differential Privacy

- Advantages
  - Differential Privacy is independent of the dataset; it is a property of the release mechanism
  - Provides strong theoretical guarantees
  - (Almost) no assumption on external knowledge

- Disadvantages
  - Sometimes requires adding too much noise;
    - Destroys utility of the data
  - Difficult to set the privacy budget ϵ

Privacy in Practice

- So far, k-anonymity and differential privacy have seen little use in practice (in progress)
  - Adding noise or modify the dataset may not be acceptable from a utility point of view

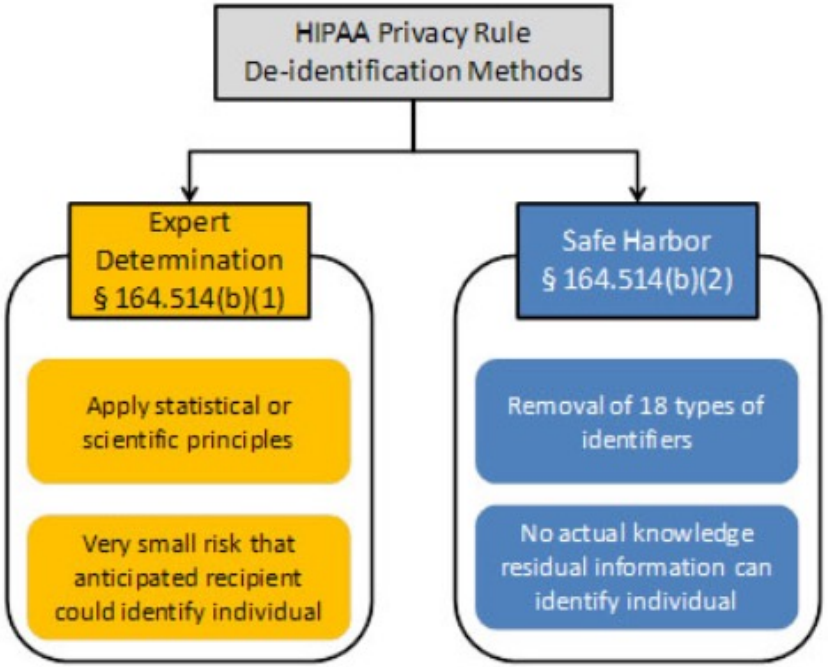
- In practice:
  - Legal considerations, e.g., HIPAA Privacy Rule
  - Data Use Agreements (DUAs)

HIPAA

- Health Insurance Portability and Accountability Act (HIPAA) 1996
  - In particular, addresses security and privacy of health data

- HIPAA Privacy Rule
  - Two options for de-identification
    1. Safe Harbor: redaction of 18 sensitive attributes
    2. Expert Determination: e.g., statistician certifies risk of re-identification is “small”

HIPAA De-Identification



[hhs.gov]

Terminology

- Protected Health Information (PHI): identifying information about
  - An individual's physical or mental health
  - An individual's provision of health care
  - E.g., laboratory report, medical bill
- Covered Entity:
  1) Health care provider
  2) Health care clearinghouse
  3) Health plan
- Standard de-identification of PHI:
  - Information is not individually identifiable
  - There is no reasonable basis to believe that re-identification can occur

Safe Harbor

(2)(i) The following identifiers of the individual or of relatives, employers, or household members of the individual, are removed:

(A) Names	
(B) All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial three digits of the ZIP code if, according to the current publicly available data from the Bureau of the Census: (1) The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; and (2) The initial three digits of a ZIP code for all such geographic units containing 20,000 or fewer people is changed to 000	
(C) All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older	
(D) Telephone numbers	(L) Vehicle identifiers and serial numbers, including license plate numbers
(E) Fax numbers	(M) Device identifiers and serial numbers
(F) Email addresses	(N) Web Universal Resource Locators (URLs)
(G) Social security numbers	(O) Internet Protocol (IP) addresses
(H) Medical record numbers	(P) Biometric identifiers, including finger and voice prints
(I) Health plan beneficiary numbers	(Q) Full-face photographs and any comparable images
(J) Account numbers	(R) Any other unique identifying number, characteristic, or code, except as permitted by paragraph (c) of this section [Paragraph (c) is presented below in the section "Re-identification"]; and
(K) Certificate/license numbers	

(ii) The covered entity does not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information.

Safe Harbor

(2)(i) The following identifiers of the individual or of relatives, employers, or household members of the individual, are removed:

(B) All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial three digits of the ZIP code if, according to the current publicly available data from the Bureau of the Census:

(C) All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older

(F) Email addresses	(N) Web Universal Resource Locators (URLs)
(G) Social security numbers	(O) Internet Protocol (IP) addresses
(H) Medical record numbers	(P) Biometric identifiers, including finger and voice prints
(I) Health plan beneficiary numbers	(Q) Full-face photographs and any comparable images
(J) Account numbers	(R) Any other unique identifying number, characteristic, or code, except as permitted by paragraph (c) of this section [Paragraph (c) is

(ii) The covered entity does not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information.

References

- [Sweeney02]: Sweeney, Latanya. "k-anonymity: A model for protecting privacy." International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10.05 (2002): 557-570.
- [NS08]: Narayanan and Shmatikov. "Robust de-anonymization of large sparse datasets." IEEE S&P 2008.
- [Dwork06]: Dwork, Cynthia. "Differential privacy." ICALP 2006.

Discussion Questions

- [Homer et al. 08] Genome-Wide Association Study (GWAS)
  - Study looks at SNPs of a population - link those to a disease
  - Two groups: control group, and disease group
  - Aggregate statistics can disclose membership
  - Attack requires genome of the target individual

1. What should be considered a privacy violation?
   - Attribute disclosure?
   - Membership disclosure?
   - something else?

   (Recall that attribute disclosure can happen even when the target individual is not in the dataset, e.g., “smoking causes cancer.”)

Discussion Questions

2. What techniques would you use to de-identify a dataset?
   - Technical (e.g., k-anonymity, differential privacy)?
   - Legal (e.g., DUAs)?
   - Both?