Security and Trust I: 5. Privacy

Dusko Pavlovic

UHM ICS 355
Fall 2014

Outline

Idea of privacy

Surveillance and sousveillance

Database privacy

Lesson

Outline

Idea of privacy

Concept

Privacy in US legal history

Privacy is deeper

Surveillance and sousveillance

Database privacy

Lesson

What is privacy?

What is privacy?

Privacy is the right to be left alone.

What does privacy have to do with security?

What does privacy have to do with security?

- Security is the adversarial *process* of defending and attacking some privately owned assets.

What does privacy have to do with security?

- Security is the adversarial *process* of defending and attacking some privately owned assets.
- Privacy is the owners' *right* to enjoy their assets with no interference from others.

What does privacy have to do with security?

- Security is the process whereby some assets are
  - made and kept private by the owners
  - reassigned to other owners or made public
- Privacy is a security requirement
  - to implement the claimed *"natural laws"*
    - "the data about me are owned by me"
    - "the sea is owned by the king"

The private vs the public

Aristotle's *Politics* (∼330 BC)

private sphere: family, home, childbirth, household

- oikos (οίϰος) ⇝ economy, economics

public sphere: city, market, war, constitutions

- polis (πόλις) ⇝ policy, politics

The private vs the public

Sophocles: The tragedy of *Antigona* (441 BC)

private sphere: family, home, childbirth, household

- Antigona's brothers Eteocles and Polyneices are on two sides in a war

public sphere: city, market, war, constitutions

- Polyneices' side loses and King Creon orders that his body be left to rot in the battlefield

The private vs the public

Antigona is torn between

private sphere: family, home, childbirth, household

- **the duty to bury her brother**

public sphere: city, market, war, constitutions

- **the duty to obey the king**

The private vs the public

Antigona is torn between

private sphere: family, home, childbirth, household

- **the duty to bury her brother**

public sphere: city, market, war, constitutions

- **the duty to obey the king**

This **tragic conflict** has pursued the mankind ever since.

Modern legal treatment of privacy

Warren and Brandeis (1890)

In very early times, the law gave a remedy only for physical interference with life and property, for trespasses *vi et armis*. Then the "right to life" served only to protect the subject from battery in its various forms; liberty meant freedom from actual restraint; and the right to property secured to the individual his lands and his cattle. Later, there came a recognition of man's spiritual nature, of his feelings and his intellect. Gradually the scope of these legal rights broadened; and now the right to life has come to mean the right to enjoy life — the right to be let alone; the right to liberty secures the exercise of extensive civil privileges; and the term "property" has grown to comprise every form of possession – intangible, as well as tangible.

Privacy is not in the US Constitution

Fourth Amendment comes close

The right of the people to be secure in their persons, houses, papers and effects against unreasonable searches and seizures shall not be violated, and no Warrants shall issue but upon probable cause, supported by Oath or affirmation, and particularly describing the place to be searched, and the persons or things to be seized.

Privacy is not in the US Constitution

Fourth Amendment context

- Trying to improve taxation on the imports in the Colonies, the Crown introduced had introduced the *writs of assistance*, which empowered officers of the Crown to search *"wherever they suspected uncustomed goods to be"* and to *"break open any receptacle or package falling under their suspecting eye"*
- Fourth Amendment curtails such sweeping searches.
- Protections of rights less tangible than *"persons, houses, papers and effects"* took a long to evolve.

New communication channels

Lous Brandeis (dissent *Olmstead v US*)

The evil incident to invasion of privacy of the telephone is far greater than that involved in tampering with the mails. Whenever a telephone line is tapped, the privacy of persons at both ends of the line is invaded, and all conversations between them upon any subject, and although proper, confidential and privileged, may be overheard. Moreover, the tapping of one man's telephone line involves the tapping of the telephone of every other person whom he may call or who may call him. As a means of espionage, writs of assistance and general warrants are but puny instruments of tyranny and oppression when compared with wire-tapping.

. . . require new privacy protections

Lous Brandeis (dissent *Olmstead v US*)

The makers of our Constitution undertook to secure conditions favorable to the pursuit of happiness. They recognized the significance of man's spiritual nature, of his feelings and his intellect. [. . .] They sought to protect Americans in their beliefs, their thoughts, their emotions and their sensations. They conferred, as against the Government, the right to be let alone — the most comprehensive of rights and the right most valued by civilized man. To protect that right, every unjustifiable intrusion by the Government upon the privacy of the individual, whatever the means employed, must be deemed a violation of the Fourth Amendment. And the use, of evidence in a criminal proceeding, of facts ascertained by such intrusion must be deemed a violation of the Fifth.

. . . and maintaining the old ones

Watergate hearings (1973)

Sen. Herman Talmadge: Do you remember when we were in law school, we studied a famous principle of law that came from England and also is well known in this country, that no matter how humble a man's cottage is, that even the King of England cannot enter without his consent.

Witness John Ehrlichman: I am afraid that has been considerably eroded over the years, has it not?

Sen. Talmadge: Down in my country we still think of it as a pretty legitimate piece of law.

Privacy goes deeper: Culture

What are the private areas?

- **home:** multi-level security
  - privacy from the outsiders
  - privacy from each other: children from parents. . .
  - public private spaces: bathrooms. . .

Privacy goes deeper: Culture

What are the private areas?

- **home:** multi-level security
  - privacy from the outsiders
  - privacy from each other: children from parents. . .
  - public private spaces: bathrooms. . .
- **body:** private areas vs public areas separated by clothes
  - sex is the realm of privacy
  - the view of the private areas can be
    - monetized (stripping, pornography)
    - owned by others (in many traditions)

Privacy goes even deeper: Biology

Cooperation vs competition

- The evolution from solitary wasps to social insects shows the function of the public sphere
  - Cooperation benefits all.
- The private assets of the dominant individuals in the hierarchical societies shows the function of the private sphere
  - Privacy benefits the winners
  - *Private vices public benefits* (B. Mandeville)

Privacy goes higher: Social technology

Data gathering and processing

- weakens the privacy of
  - citizens
  - consumers
- strengthens the privacy of
  - governments
  - industries

Outline

Idea of privacy

Surveillance and sousveillance

- Surveillance and copy protections
- Sousveillance

Database privacy

Lesson

Privacy and ownership

Private data and private property are closely related:

- individual privacy vs surveillance
  - HIPAA, Experian, Doubleclick, Echelon
- copy protections vs file sharing
  - DRM, DMCA, thepiratebay

Surveillance and file sharing

Data security problem

Protect data confidentiality and authenticity

Surveillance and file sharing

Data security problem

Protect data confidentiality and authenticity

Data security areas

- individual privacy
  - identity
  - behavior
    - shopping
    - networking

⇓

- Surveillance attacks

- copy protections
  - patents
  - entertainment
    - music
    - film

⇓

- DRM defenses

Surveillance and file sharing

If the data security tasks are split into

- **privacy** when the data are *about* the subject
- **copy protections** when the data are *owned* by the subject

then the corresponding attack models are

- **surveillance** against privacy
- **file sharing** against copy protections

Surveillance and Digital Rights Management

Jonathan Zittrain (2000), Larry Lessig (2002)

- The tasks of securing
  - data privacy
  - intellectual property

  give rise to **the same security problem**:
  - control the data flows in digital networks

Surveillance and Digital Rights Management

Jonathan Zittrain (2000), Larry Lessig (2002)

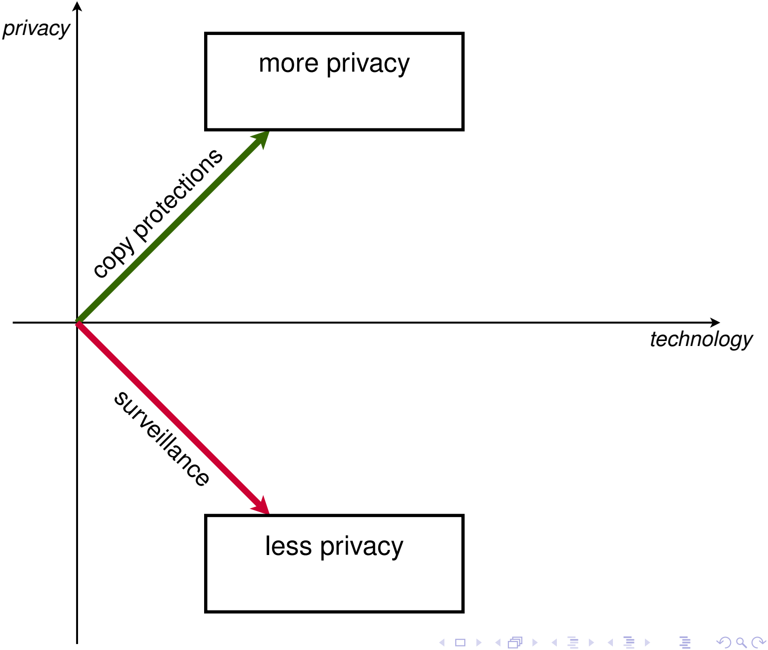
- The tasks of securing
  - data privacy
  - intellectual property

  give rise to **the same security problem**:
  - control the data flows in digital networks

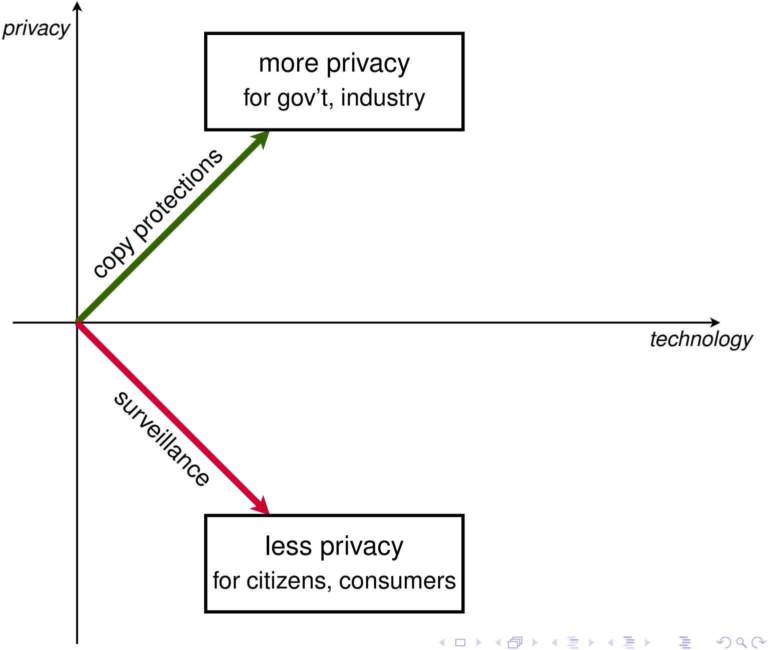
- The technologies developed for these tasks
  - surveillance
  - copy protections

  lead to **the opposite solutions**:
  - weakening privacy
  - strengthening intellectual property

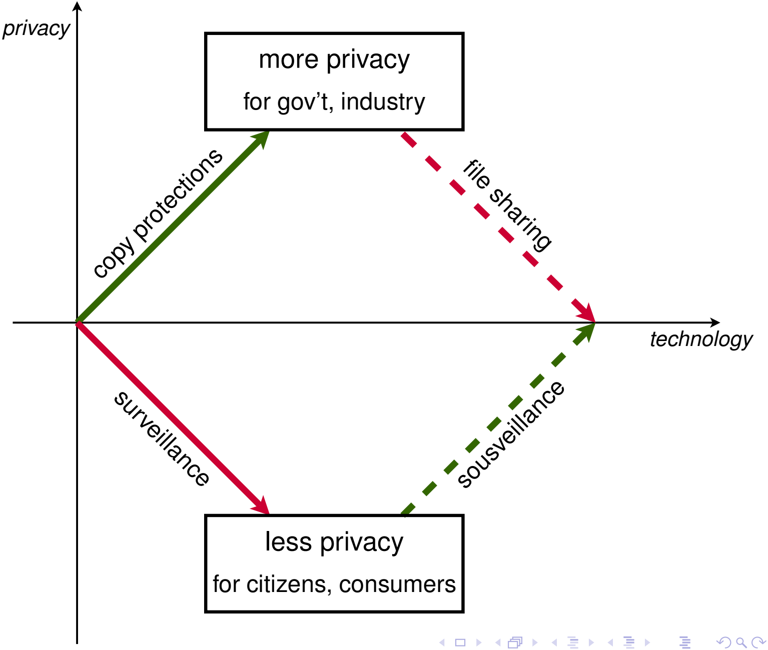
Privacy and technology



Privacy and technology



Privacy and technology



privacy

more privacy

for gov't, industry

copy protections

file sharing

technology

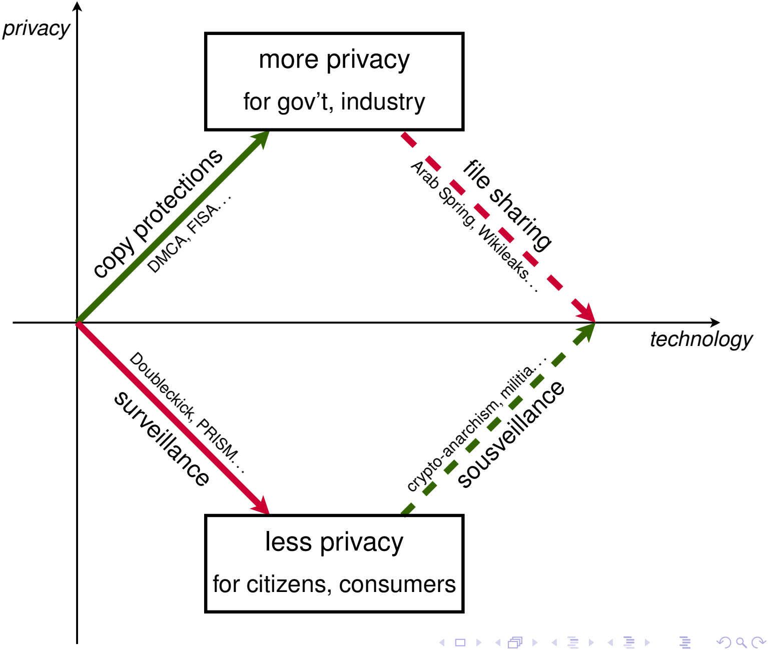
surveillance

sousveillance

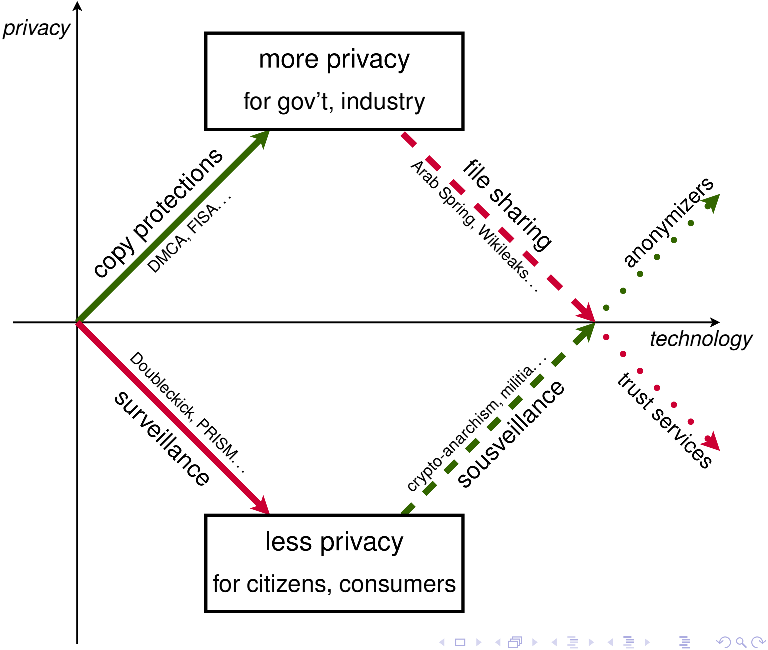
less privacy

for citizens, consumers

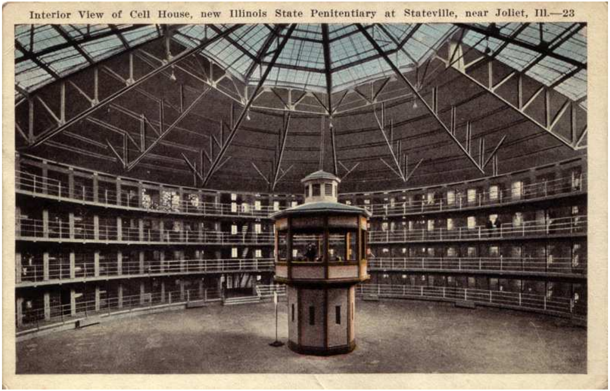
Privacy and technology



Privacy and technology

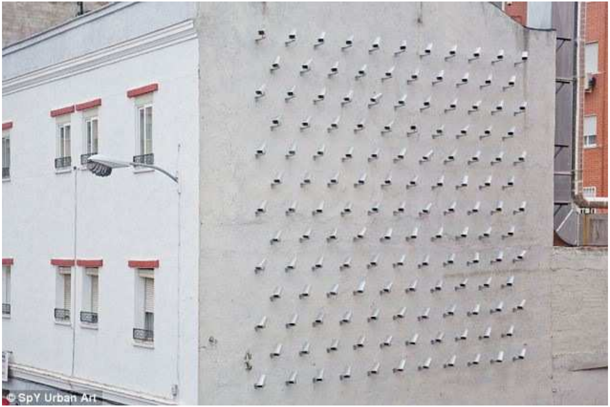


Surveillance technique: Panopticon


Interior View of Cell House, new Illinois State Penitentiary at Stateville, near Joliet, Ill.—23

Jeremy Bentham: Architecture to eliminate privacy

Surveillance technique: CCTV



© SpY Urban Art

There is one CCTV camera on every 11 citizens of UK

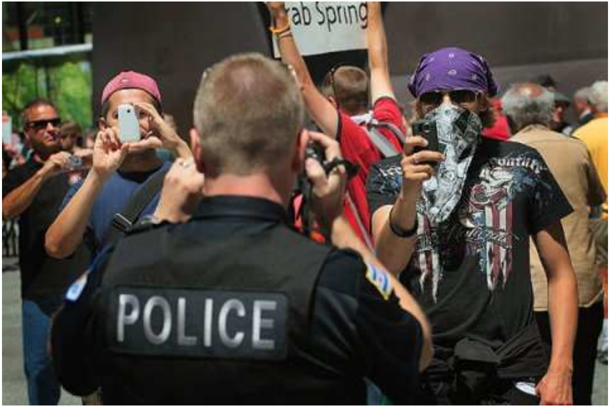
Surveillance and security industry

Security industry is a powerful interest group in the US

Surveillance: Positive feedback

. . . crime → surveillance → enforcement → crime. . .

Surveillance: Negative feedback



surveillance ↔ sousveillance

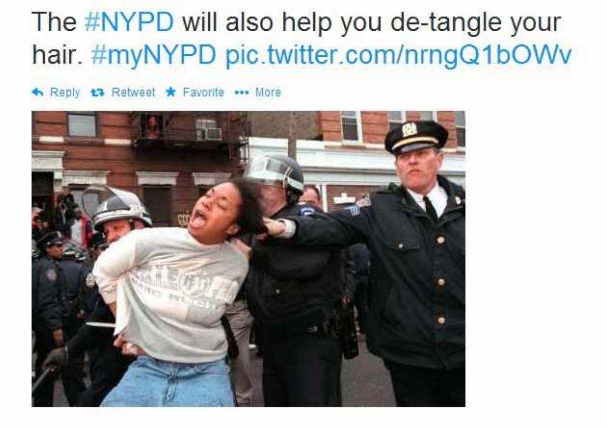
Sousveillance: Countersurveillance

Social networking enables surveillance from below

Sousveillance: "Arab Spring"

Public information may or may not bring real power

Sousveillance: #myNYPD, #myLAPD. . .



The #NYPD will also help you de-tangle your hair. #myNYPD pic.twitter.com/nrngQ1bOWv

New technologies erode the privacy of public service

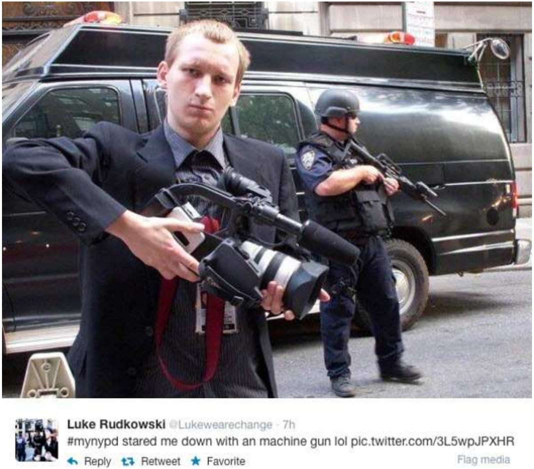
Balance of veillances

New technologies make public service more public

Balance of veillances

New technologies make private life less private

Balance of veillances



Luke Rudkowski @Lukewearechange · 7h

#mynypd stared me down with an machine gun lol pic.twitter.com/3L5wpJPXHR

New technologies do not transfer power, just information

Balance of powers

Private life and political power can be hard to separate

Outline

Idea of privacy

Surveillance and sousveillance

Database privacy

Lesson

Data privacy

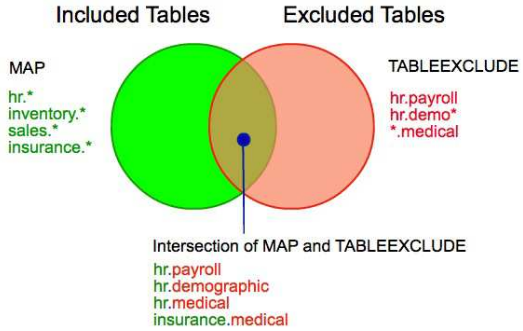
The life of data consists of

- data gathering (veillance)
- data storage and release (databases)
- data processing (mining and classification)

Data privacy

Surveillance alone does not kill privacy

Data privacy



Database search kills privacy

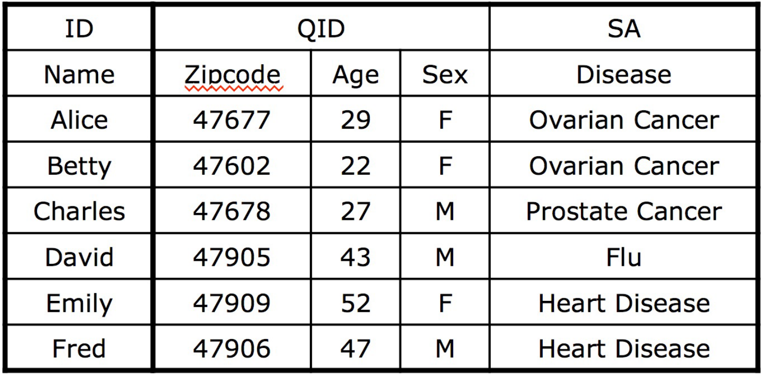
Problem of anonymizing databases

Statistical databases need to be anonymized

- Data are often used to calculate sums, averages, statistics
  - voting, market research, science, medicine. . .
- *Statistical database* is a database released for statistical research
  - to calculate averages, correlations. . .
- If a statistical database contains private data, then it needs to be ***anonymized***

Example

Medical database

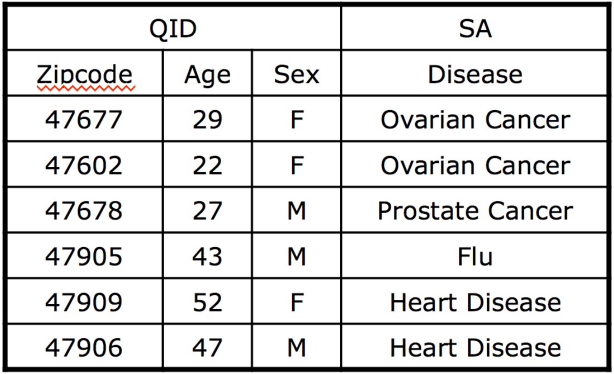


ID	QID			SA
Name	Zipcode	Age	Sex	Disease
Alice	47677	29	F	Ovarian Cancer
Betty	47602	22	F	Ovarian Cancer
Charles	47678	27	M	Prostate Cancer
David	47905	43	M	Flu
Emily	47909	52	F	Heart Disease
Fred	47906	47	M	Heart Disease

Data contain identifiers (ID) and sensitive attributes (SA).

Example

Medical database: Simple anonimization

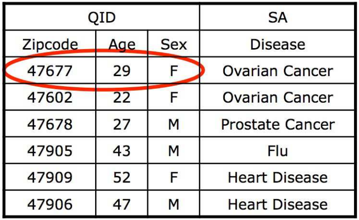


QID			SA
Zipcode	Age	Sex	Disease
47677	29	F	Ovarian Cancer
47602	22	F	Ovarian Cancer
47678	27	M	Prostate Cancer
47905	43	M	Flu
47909	52	F	Heart Disease
47906	47	M	Heart Disease

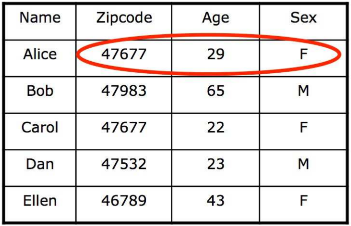
To maintain confidentiality of SA, omit the IDs.

Example

Medical database linked with Voter Register



QID			SA
Zipcode	Age	Sex	Disease
47677	29	F	Ovarian Cancer
47602	22	F	Ovarian Cancer
47678	27	M	Prostate Cancer
47905	43	M	Flu
47909	52	F	Heart Disease
47906	47	M	Heart Disease

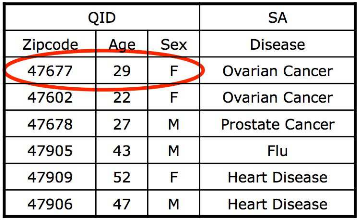


Name	Zipcode	Age	Sex
Alice	47677	29	F
Bob	47983	65	M
Carol	47677	22	F
Dan	47532	23	M
Ellen	46789	43	F

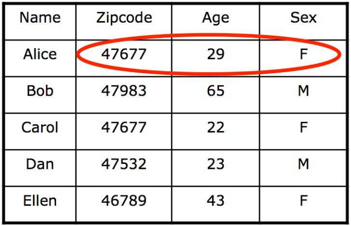
Linking databases allows re-identification

Example

Medical database linked with Voter Register



QID			SA
Zipcode	Age	Sex	Disease
47677	29	F	Ovarian Cancer
47602	22	F	Ovarian Cancer
47678	27	M	Prostate Cancer
47905	43	M	Flu
47909	52	F	Heart Disease
47906	47	M	Heart Disease

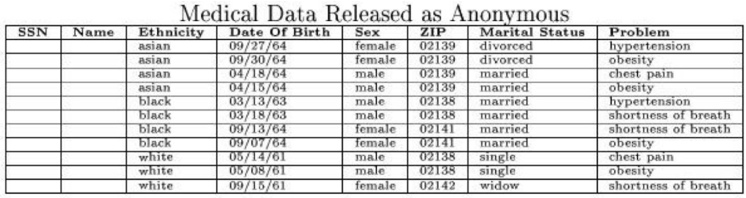


Name	Zipcode	Age	Sex
Alice	47677	29	F
Bob	47983	65	M
Carol	47677	22	F
Dan	47532	23	M
Ellen	46789	43	F

Linking databases allows re-identification
whenever quasi-identifier (QID) corresponds to unique ID

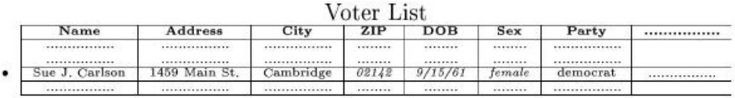
Such examples are real

Medical database linked with Voter Register



Medical Data Released as Anonymous

SSN	Name	Ethnicity	Date Of Birth	Sex	ZIP	Marital Status	Problem
		asian	09/27/64	female	02139	divorced	hypertension
		asian	09/30/64	female	02139	divorced	obesity
		asian	04/18/64	male	02139	married	chest pain
		asian	04/15/64	male	02139	married	obesity
		black	03/13/63	male	02138	married	hypertension
		black	03/18/63	male	02138	married	shortness of breath
		black	09/13/64	female	02141	married	shortness of breath
		black	09/07/64	female	02141	married	obesity
		white	05/14/61	male	02138	single	chest pain
		white	05/08/61	male	02138	single	obesity
		white	09/15/61	female	02142	widow	shortness of breath



Voter List

Name	Address	City	ZIP	DOB	Sex	Party	....
....	....	....	....	....	....	....	
....	....	....	....	....	....	....	
Sue J. Carlson	1459 Main St.	Cambridge	*02142*	*9/15/61*	*female*	democrat	....
....	....	....	....	....	....	....	

Medical record of the Governor of Massachusetts identified

Task

Find methods to prevent this.

When is a database anonymized?

Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule

Under the safe harbor method, covered entities must remove all of a list of 18 enumerated identifiers and have no actual knowledge that the information remaining could be used, alone or in combination, to identify a subject of the information.

When is a database anonymized?

Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule

The identifiers that must be removed include direct identifiers, such as name, street address, social security number, as well as other identifiers, such as birth date, admission and discharge dates, and five- digit zip code. The safe harbor requires removal of geographic subdivisions smaller than a State, except for the initial three digits of a zip code if the geographic unit formed by combining all zip codes with the same initial three digits contains more than 20,000 people. In addition, age, if less than 90, gender, ethnicity, and other demographic information not listed may remain in the information. The safe harbor is intended to provide covered entities with a simple, definitive method that does not require much judgment by the covered entity to determine if the information is adequately de-identified.

When is a database anonymized?

Dalenius Desideratum

All sensitive data about an individual i that can be learned from the database D can also be learned without access to D.

Tore Dalenius, 1977

When is a database anonymized?

Dalenius Desideratum

No sensitive data about an individual *I* should be learnable from the database *D* that cannot be learned without access to *D*.

Tore Dalenius, 1977

When is a database anonymized?

Trouble

Suppose that

- risk of heart attack is accepted as sensitive attribute
- database *D* suggests a correlation between heart attack and eating a lot of chocolate
- it is publicly known that Dusko eats a lot of chocolate

When is a database anonymized?

Trouble

Suppose that

- risk of heart attack is accepted as sensitive attribute
- database *D* suggests a correlation between heart attack and eating a lot of chocolate
- it is publicly known that Dusko eats a lot of chocolate

Database *D* thus discloses Dusko's private data whether his record is included in it or not.

Model database

Definition

Given the sets

- $\mathcal{R}$ of *records*,
- $\mathcal{A}$ of *attributes*
- V_a of *values* for each $a \in \mathcal{A}$

a database is a matrix

$$D \quad : \quad \mathcal{R} \times \mathcal{A} \to V$$

where $V = \bigcup_{a \in \mathcal{A}} V_a$ and $D(r, a) \in V_a$ for all $r \in \mathcal{R}$ and $a \in \mathcal{A}$.

Model database

Dictionary for database books and papers

- matrix = *table*.
- row = *tuple* (of data in a record)
- column = *attribute* (data for an attribute)

Model data collection and processing

Definition

Data are collected from a set of *entities* $\mathcal{E}$.

- *Data gathering* is a map $R : \mathcal{E} \to \mathcal{R}$, so that $D_{R(e)}$ is the tuple of the data corresponding to the entity $e \in \mathcal{E}$.
- *Data identification* is a map $E : \mathcal{R} \to \mathcal{E}$, such that $E(R(e)) = e$.

Model data collection and processing

Definition

An *identifier* (ID) is an attribute $i \in \mathcal{A}$ that uniquely determines any entity.

Model data collection and processing

Definition

An *identifier* (ID) is an attribute $i \in \mathcal{A}$ that uniquely determines any entity.

More precisely, there is $f : V_i \to \mathcal{E}$ such that for all $e \in \mathcal{E}$ holds

$$f(D^i_{R(e)}) \quad = \quad e$$

Model data collection and processing

Definition

A *quasi-identifier* (QID) is a set of attributes $Q \subseteq \mathcal{A}$ that uniquely determine some entities.

Model data collection and processing

Definition

A *quasi-identifier* (QID) is a set of attributes $\mathcal{Q} \subseteq \mathcal{A}$ that uniquely determine some entities. More precisely, there is a partial function $f : \prod_{i \in \mathcal{Q}} V_i \rightharpoonup \mathcal{E}$ such that for some $e \in \mathcal{E}$ holds

$$f(D^{\mathcal{Q}}_{R(e)}) \quad = \quad e$$

where $D^{\mathcal{Q}}_{R(e)}$ is a $\mathcal{Q}$-tuple of attributes in the database $\mathcal{D}$

Model data privacy

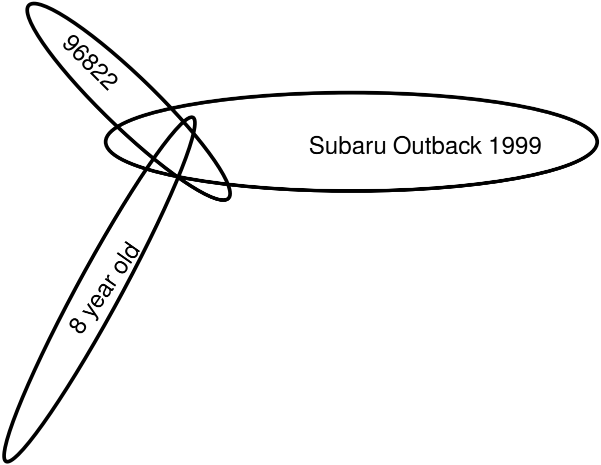
Definition

A database D satisfies the *k-anonymity requirement* if for every quasi-identifier Q and every Q-tuple of values D^Q there are

- either at least k records with the same value D^Q
- or no such records, i.e. D^Q does not come about in D.

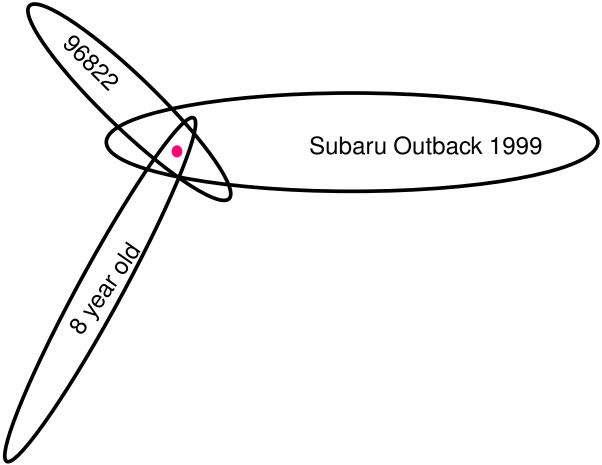
Idea of *k*-anonymity

QID $\langle ZIP, car, child\rangle$



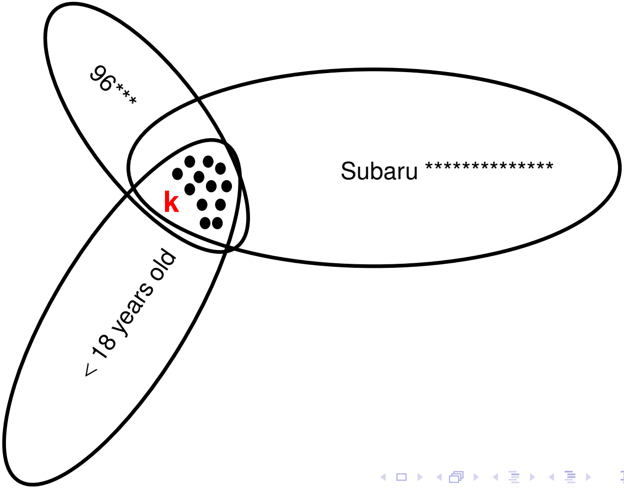
Idea of *k*-anonymity

QID $\langle ZIP, car, child \rangle$



Idea of *k*-anonymity

QID $\langle ZIP, car, child \rangle$ — k-anonymized



Methods to achieve *k*-anonymity

- **Generalization:** replace the precise QID values with a more general value
  - when the precise values together average out to the general value
- **Suppression:** suppress the records containing the "outlier" values
  - generalizing the values far from other values would cause the distortion of the average and statistics

Problems with *k*-anonymity

- **Lack of diversity:** If the same SA value occurs in more than *k* records, then *k*-anonymity does not conceal it
  - Database may be *k*-anonymous, and disclose SA.
- **Background information:** General anonymized data may disclose individual SA combined with the background information about an individual.
  - The data relating smoking and cancer from database *D*, together with the knowledge that Bob smokes, link Bob with the SA of cancer risk — even if Bob does not occur in *D*.

Background information is a **false problem**

Fact

Anonymizing database *D* cannot eliminate the information available outside *D*.

Background information is a **false problem**

Fact

Anonymizing database *D* cannot eliminate the information available outside *D*.

Consequence

I must accept that a database *D* may disclose some sensitive information about me to those who know me — even if I do not occur in *D*.

Idea

Differential Privacy

All sensitive data about an individual i that
can be learned from the database D with a record $r(i)$
can also be learned from the database D' where $r(i)$ is
replaced by any r'.

Cynthia Dwork 2008

Looks like a small step?

Dalenius Desideratum

All sensitive data about an individual *i* that
can be learned from the database *D*
can also be learned without access to *D*.

Tore Dalenius, 1977

No, it is a big step

The devil is in the details

Differential privacy

- is a requirement on the disclosure algorithm F, not on the database D
- implements the indistinguishability of databases D and D' in terms of an equivalence kernel
  - We used equivalence kernels to quantify flow security in Lecture 4.
  - Differential privacy requires that the flow leakage of individual information is negligible.

Differential privacy

Definition

Let

- $\mathcal{D}$ be a family of databases,
- $\mathcal{P} \subseteq \sum_{a\in\mathcal{A}} V_a$ a family of properties (viewed as sets of values in some attributes), and
- $\varepsilon > 0$ a real number.

A disclosure algorithm $F : \mathcal{D} \to \mathcal{P}$ is ε*-differentially private* if for every property $Y \in \mathcal{P}$ holds

$$\left| \frac{\Pr(F(D) \in Y)}{\Pr(F(D') \in Y)} \right| \quad \leq \quad e^{-\varepsilon}$$

for any pair $D, D' \in \mathcal{D}$ which differ in at most one record.

Differential privacy

Remark

Recall that the normalized ratio is was defined by

$$\vdash\!\frac{x}{y}\!\dashv \quad = \quad \begin{cases} \frac{x}{y} & \text{if } x \leq y \\ \frac{y}{x} & \text{if } x > y \end{cases}$$

so that

$$\vdash\!\frac{\Pr(F(D) \in Y)}{\Pr(F(D') \in Y)}\!\dashv \quad \leq \quad e^{\varepsilon}$$

is thus equivalent with

$$\left|\log \Pr(F(D) \in Y) - \log \Pr(F(D') \in Y)\right| \quad \leq \quad \varepsilon$$

Differential privacy

Explanation

The difference between the attacker's information from D and from D' is indistinguishable in the same sense in which his prior and posterior beliefs were indistinguishable when extracting information from probabilistic channels in Lecture 4.

Methods to achieve differential privacy

Add noise at the various points of the disclosure process:

- output perturbation
- input perturbation
- intermediate values

Output perturbation method

Theorem

Let $f : \mathcal{D} \to \mathcal{P}$ be a feasible disclosure algorithm. Then

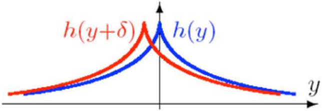
$$F(x) \quad = \quad f(x) + Lap\left(\frac{GS_f}{\varepsilon}\right)$$

is ε-differentially private, where

- $GS_f = \bigwedge_{x,x'} \|f(x) - f(x')\|$ is the *global sensitivity*
- $Lap(\lambda)$ is the Laplace distribution

Output perturbation method

Proof idea



Sliding property of Lap$\left(\frac{\mathsf{GS}_f}{\varepsilon}\right)$: $\frac{h(y)}{h(y+\delta)} \leq e^{\varepsilon \cdot \frac{\|\delta\|}{\mathsf{GS}_f}}$ for all y, δ

because the density of $Lap(\lambda)$ is

$$h(y) \quad \propto \quad e^{-\frac{\|y\|}{\lambda}}$$

Outline

Idea of privacy

Surveillance and sousveillance

Database privacy

Lesson

What did we learn?

- Privacy is the right to be left alone.
- The balance of the public sphere and the private sphere is a balance of political powers.
- The same technologies provide more privacy for those in power and less privacy for those under control.
- The new technologies facilitate both surveillance from above and sousveillance from below.
- Techniques to assure database privacy have a significant social impact.