Privacy Concerns: The Clash between Technological Capabilities and Societal Expectations

Ernst L. Leiss
Department of Computer Science
University of Houston
coscel@cs.uh.edu

Partially funded under NSF Grant #1241772

Any opinions, findings, conclusions, or recommendations expressed herein are those of the author and do not reflect the views of the National Science Foundation

Updated 14 Nov. 2014

1. **Introduction: Societal Expectations**
2. **Statistical databases: Inference control**
3. **Data aggregation**
4. **Monitoring the location of an individual: Cell phones, tracking devices for vehicles, devices that read license plates of vehicles and monitor their location, face recognition software for public monitoring cameras**
5. **Monitoring the behavior of an individual: Cars with devices that record speed, acceleration and deceleration, g-forces (in turns), recording websites visited**
6. **Monitoring communications: Cell phones vs. landlines (US), monitoring under a court order, monitoring international communications (Echelon)**
7. **Encryption: Access to encrypted communication under a court order, requiring the divulgence of keys and passwords (self-incrimination), treating encryption algorithms as munitions (US) or restricting/forbidding its use**
8. **SMS, texting, etc.: For business purposes, the lack of a central storage and monitoring unit may be undesirable**
9. **Repurposing data sets collected for a different purpose**
10. **The use of watermarking to trace digital content**
11. **Implanted RFIDs for people?**
12. **The use of an individual's fully sequenced genome**
13. **Conclusion**

1. Introduction: Societal Expectations

Behavior: Law, ethics
For individuals, organizations, and government

Laws should reflect ethics

Technology: Unexpected and unknown capabilities
May clash with a society's expectations and ethics

The expectation of privacy: may be contradicted by new technological capabilities

Legal landscape: US vs. EU

This talk is concerned with the tension between what we <u>can</u> do and what we <u>should</u> do

2. Statistical databases: Inference control

The notion of a statistical database

- **Access to individual entries vs. access to statistics**

Confidential data, legal requirements to maintain confidentiality:

- **The use of statistical queries must guarantee confidentiality**

In practice, extremely difficult to achieve

Extensive literature, for many approaches, restrictions, and models of statistical databases

Example

Assume ordinary SQL queries

- **Each query involves a set of elements**
- **Typical: Median, average, sum**
- **Confidential information in statistical queries is numerical**
- **Assume query type SUM: A query defines a set of elements and returns the sum of the confidential information associated with the elements in the set**

Can do or (union of underlying set) and not (complement of set)

Restriction: A query q is legal iff h <= NU(q) <= N-h

- **N is the total number of elements in the database, h is an arbitrary value,**
- **NU(q) is the number of elements in the set underlying the query q**

General tracker: Any query GT s. t. 2h <= NU(GT) <= N-2h

Illegal query q_{bad}:

x := SUM(GT) + SUM(notGT)

$SUM(q_{bad}) = SUM(q_{bad}$ or GT) + SUM(q_{bad} or not GT) – x too small

$SUM(q_{bad}) = 2x – [SUM(not\ q_{bad}$ or GT) + SUM(not q_{bad} or not GT)] too large

Upshot

- Impossible to maintain the confidentiality of the information of individual entries

Similar results hold for virtually all type of restrictions that are imposed on the queries

Randomizing is a possible solution, except the responses are falsified (slightly)

3. Data aggregation

Two entirely innocuous databases may be combined to identify uniquely an individual

Example:

US: DoB; 5-digit zip code

330M; 100 000 zip codes: on average 3 300 inhabitants per zip

365 x 80 birth dates: on average 12 000 people with same DoB

Combined: DoB plus zip code identifies uniquely many, if not most individuals: 100 000 x 365 x 80 >> 330M

4. Monitoring the location of an individual

Cell phones: Store and transmit location information, service providers may provide this information

Tracking devices for vehicles: Can be used to locate individuals, may be attached without the owner's knowledge

Devices that read license plates of vehicles and monitor their location
Law enforcement, traffic restrictions (Inner City), toll booths

Face recognition software for public monitoring cameras

Ownership of data collected

5. Monitoring the behavior of an individual

Cars with devices that record driving behavior

speed, acceleration and deceleration, g-forces (in turns)

Recording websites visited

websites, time, duration

Point-of-Sales information

acquisition of items, matching credit card information with purchases

Ownership of data collected

6. Monitoring communications

Cell phones vs. landlines (US)

"Americans' Cellphones Targeted in Secret U.S. Spy Program"

(Nov 14, 2014, WSJ) Devices on planes mimic cellphone towers to target criminals' phones but also gather information from thousands of other phones

Ownership of communication systems: Employers monitoring employees' e-mail and surfing/downloading behavior

Monitoring under a court order

Monitoring international communications (Echelon)

7. Encryption

Encryption is dual-use: Civilian and government (law-enforcement, military)

Access to encrypted communication under a court order

Requiring the divulgence of keys and passwords (self-incrimination)

Treating encryption algorithms as munitions (US) or restricting/forbidding its use

8. SMS, texting, etc.

Ubiquitous technology

Private use

Business use: lack of central storage and monitoring unit undesirable

Lack of control

9. Repurposing Data Sets

In much scientific research, data sets are collected.
Permissions to use such data sets must often be secured.
These permissions often define specific purposes of study and analysis.

It is frequently tempting to reuse such a data set for a different study or analysis.
Different data sets may also be combined with the same study or analysis objective.

It is necessary to obtain permissions for the new studies or analyses?

EU's privacy directive vs. US sectoral legislation (related to medical, financial, or student data)

Example: Human subjects. IRBs. Medical data sets.

10. Watermarking to trace digital content

Digital watermarks

Tag digital objects imperceptibly (to the human senses)

Video and audio

Can be used to identify uniquely copies of digital objects

Precedent: The Oscars

11. Implanted RFIDs for people?

Exist already for pets

Easily applied to people (e. g., children)

Monitoring requires extensive infrastructure (software dependent)

Orwellian society

12. The use of an individual's fully sequenced genome

An individual's fully sequenced genome costs now under US$1000 (2014)

Drop in cost largely due to more efficient software (and economy of scale)

Is it desirable for many people to know their full genome?

- **to predict disease susceptibility [Huntingdon's – no cure!]**
- **for genetic screening, e. g., for jobs and for insurance purposes**
- **for fetal testing, including to determine termination of pregnancies**
- **for genetic counseling**

13. Conclusion

Raised issues of clashes between technological capabilities and ethics

Most violate our expectation of privacy, but not all (criminals encrypting their communications)

The problem of quality of information (e. g., toll roads vs. DMV info)

The law lags seriously behind the technology

As computer scientists, we have a greater responsibility than ordinary citizens because we help create the technology