Access Control and Privacy Policies (8)

Email: christian.urban at kcl.ac.uk
Office: S1.27 (1st floor Strand Building)
Slides: KEATS (also homework is there)

Man-in-the-Middle

Facebook Privacy

Privacy, Anonymity et al

Some terminology:

- secrecy is the mechanism used to limit the number of principals with access to information (eg, cryptography or access controls)
- confidentiality is the obligation to protect the secrets of other people or organizations (secrecy for the benefit of an organisation)
- anonymity is the ability to leave no evidence of an activity (eg, sharing a secret)
- privacy is the ability or right to protect your personal secrets (secrecy for the benefit of an individual)

Privacy vs Anonymity

- everybody agrees that anonymity has its uses (e.g., voting, whistleblowers, peer-review)

Privacy vs Anonymity

- everybody agrees that anonymity has its uses (e.g., voting, whistleblowers, peer-review)

But privacy?

"You have zero privacy anyway. Get over it."
Scott Mcnealy (CEO of Sun)

If you have nothing to hide, you have nothing to fear.

Privacy

private data can be often used against me

- if my location data becomes public, thieves will switch off their phones and help themselves in my home
- if supermarkets can build a profile of what I buy, they can use it to their advantage (banks - mortgages)
- my employer might not like my opinions

Privacy

private data can be often used against me

- if my location data becomes public, thieves will switch off their phones and help themselves in my home
- if supermarkets can build a profile of what I buy, they can use it to their advantage (banks - mortgages)
- my employer might not like my opinions

- one the other hand, Freedom-of-Information Act
- medical data should be private, but medical research needs data

Privacy Problems

- Apple takes note of every dictation (send over the Internet to Apple)
- markets often only work, if data is restricted (to build trust)
- Social network can reveal data about you
- have you tried the collusion extension f FireFox?
- I do use Dropbox, store cards

- next week: anonymising data

Gattaca (1997)

Privacy

- we do want that government data is made public (free maps for example)
- we do not want that medical data becomes public (similarly tax data, school records, job offers)

- personal information can potentially lead to fraud (identity theft)

Privacy

- we do want that government data is made public (free maps for example)
- we do not want that medical data becomes public (similarly tax data, school records, job offers)

- personal information can potentially lead to fraud (identity theft)

"The reality":

- London Health Programmes lost in June last year unencrypted details of more than 8 million people (no names, but postcodes and details such as gender, age and ethnic origin)

Privacy

- we do want that government data is made public (free maps for example)
- we do not want that medical data becomes public (similarly tax data, school records, job offers)

- personal information can potentially lead to fraud (identity theft)

"The reality":

- also in June last year, Sony got hacked: over 1M users' personal information, including passwords, email addresses, home addresses, dates of birth, and all Sony opt-in data associated with their accounts.

Privacy and Big Data

Selected sources of "Big Data":

- Facebook
  - 40+ Billion photos (100 PB)
  - 6 Billion messages daily (5 - 10 TB)
  - 900 Million users
- Common Crawl
  - covers 3.8 Billion webpages (2012 dataset)
  - 50 TB of data
- Google
  - 20 PB daily (2008)
- Twitter
  - 7 Million users in the UK
  - a company called Datasift is allowed to mine all tweets since 2010
  - they charge 10k per month for other companies to target advertisement

Privacy and Big Data

Selected sources of "Big Data":

- Facebook
  - 40+ Billion photos (100 PB)
  - 6 Billion messages daily (5 - 10 TB)
  - 900 Million users
- Common Crawl
  - covers 3.8 Billion webpages (2012 dataset)
  - 50 TB of data
- Google
  - 20 PB daily (2008)
- Twitter
  - 7 Million users in the UK
  - a company called Datasift is allowed to mine all tweets since 2010
  - they charge 10k per month for other companies to target advertisement

Cookies...

"We have published a new cookie policy. It explains what cookies are and how we use them on our site. To learn more about cookies and their benefits, please view our cookie policy.

If you'd like to disable cookies on this device, please view our information pages on 'How to manage cookies'. Please be aware that parts of the site will not function correctly if you disable cookies.

By closing this message, you consent to our use of cookies on this device in accordance with our cookie policy unless you have disabled them."

Scare Tactics

The actual policy reads:

"As we explain in our Cookie Policy, cookies help you to get the most out of our websites.

If you do disable our cookies you may find that certain sections of our website do not work. For example, you may have difficulties logging in or viewing articles."

Netflix Prize

Anonymity is necessary for privacy, but not enough!

- Netflix offered in 2006 (and every year until 2010) a 1 Mio $ prize for improving their movie rating algorithm
- dataset contained 10% of all Netflix users (appr. 500K)
- names were removed, but included numerical ratings as well as times of rating
- some information was perturbed (i.e., slightly modified)

All OK?

Re-identification Attack

Two researchers analysed the data:

- with 8 ratings (2 of them can be wrong) and corresponding dates that can have a margin 14-day error, 98% of the records can be identified
- for 68% only two ratings and dates are sufficient (for movie ratings outside the top 500)

Re-identification Attack

Two researchers analysed the data:

- with 8 ratings (2 of them can be wrong) and corresponding dates that can have a margin 14-day error, 98% of the records can be identified
- for 68% only two ratings and dates are sufficient (for movie ratings outside the top 500)

- they took 50 samples from IMDb (where people can reveal their identity)
- 2 of them uniquely identified entries in the Netflix database (either by movie rating or by dates)

- Birth data, postcode and gender (unique for 87% of the US population)
- Preferences in movies (99% of 500K for 8 ratings)

Therefore best practices / or even law (HIPAA, EU):

- only year dates (age group for 90 years or over),
- no postcodes (sector data is OK, similarly in the US)
  no names, addresses, account numbers, licence plates
- disclosure information needs to be retained for 5 years

How to Safely Disclose Information?

- Is it possible to re-identify data later, if more data is released.

- Not even releasing only aggregate information prevents re-identification attacks. (GWAS was a public database of gene-frequency studies linked to diseases; you only needed partial DNA information in order to identify whether an individual was part of the study — DB closed in 2008)

Differential Privacy

User

tell me $f(x)$ $\Rightarrow$

$\Leftarrow f(x) + \text{noise}$

Database

$x_1, \ldots, x_n$

- $f(x)$ can be released, if f is insensitive to individual entries $x_1, \ldots, x_n$
- Intuition: whatever is learned from the dataset would be learned regardless of whether x_i participates

Differential Privacy

User

tell me $f(x) \Rightarrow$

$\Leftarrow f(x) + \text{noise}$

Database

$x_1, \ldots, x_n$

- $f(x)$ can be released, if f is insensitive to individual entries $x_1, \ldots, x_n$
- Intuition: whatever is learned from the dataset would be learned regardless of whether x_i participates

- Noised needed in order to prevent queries:
  Christian's salary =
  $$\sum \text{all staff} - \sum \text{all staff} \setminus \text{Christian}$$

Adding Noise

Adding noise is not as trivial as one would wish:

- If I ask how many of three have seen the Gangnam video and get a result as follows

Alice	yes
Bob	no
Charlie	yes

  then I have to add a noise of 1. So answers would be in the range of 1 to 3

- But if I ask five questions for all the dataset (has seen Gangnam video, is male, below 30, ...), then one individual can change the dataset by 5

Tor

- initially developed by US Navy Labs, but then opened up to the world
- network of proxy nodes
- a Tor client establishes a “random” path to the destination server (you cannot trace back where the information came from)

Tor

- initially developed by US Navy Labs, but then opened up to the world
- network of proxy nodes
- a Tor client establishes a "random" path to the destination server (you cannot trace back where the information came from)

- malicious exit node attack: someone set up 5 Tor exit nodes and monitored the traffic:
  - a number of logons and passwords used by embassies (Usbekistan 's1e7u0l7c', while Tunesia 'Tunesia' and India '1234')

Tor

- initially developed by US Navy Labs, but then opened up to the world
- network of proxy nodes
- a Tor client establishes a "random" path to the destination server (you cannot trace back where the information came from)

- bad apple attack: if you have one insecure application, your IP can be tracked through Tor
  - background: 40% of traffic on Tor is generated by BitTorrent

Skype

- Skype used to be known as a secure online communication (encryption cannot be disabled), but ...
- it is impossible to verify whether crypto algorithms are correctly used, or whether there are backdoors.
- recently someone found out that you can reset the password of somebody else's account, only knowing their email address (needed to suspended the password reset feature temporarily)

Take Home Point

According to Ross Anderson:

- Privacy in a big hospital is just about doable.
- How do you enforce privacy in something as big as Google or complex as Facebook? No body knows.

Similarly, big databases imposed by government