Access Control and Privacy Policies (7)

Ch Email: christian.urban at kcl.ac.uk
Office: S1.27 (1st floor Strand Building)

Slides: KEATS (also homework is there)

Facebook Privacy

- Who has a Facebook account?

Facebook Privacy

- Who has a Facebook account?
- Who keeps the list of friends private?

Facebook Privacy

- Who has a Facebook account?
- Who keeps the list of friends private?
- Who knows that this is completely pointless? (at least at the end of 2013)

Facebook Privacy

- Who has a Facebook account?
- Who keeps the list of friends private?
- Who knows that this is completely pointless? (at least at the end of 2013)

Create a fake account. Send a friend-request. Facebook answers with "People you may know" feature. Conveniently, it has also a "see all" button.

Facebook Privacy

- Who has a Facebook account?
- Who keeps the list of friends private?
- Who knows that this is completely pointless? (at least at the end of 2013)

"Our policies explain that changing the visibility of people on your friend list controls how they appear on your Timeline, and that your friends may be visible on other parts of the site, such as in News Feed, Search and on other people's Timelines. This behavior is something we'll continue to evaluate to make sure we're providing clarity."

UCAS

"The Universities and Colleges Admissions Service received more than £12m last year in return for sending targeted advertising to subscribers as young as 16.

The service, which controls admissions to UK universities and attracts 700,000 new applicants each year, sells the access via its commercial arm, Ucas Media.

Vodafone, O2, Microsoft and the private university accommodation provider Pure Student Living are among those who have marketed through Ucas, which offers access to over a million student email addresses...

Applicants can opt out of receiving direct marketing, but only at the cost of missing out on education and careers mailings as well."

The Guardian, 12 March 2014

Verizon

1. Device sends an HTTP request.

2. Verizon injects an HTTP header ("X-UIDH"). It's a temporary ID, hashed or HMACed with a key.

3. Destination website (or third party) receives HTTP request with injected header.

4. Website directs request to advertising exchange.

5. Advertisers on the exchange can issue a paid API call to Verizon.

6. Verizon maps the header to a temporary ID, and returns the ID and/or advertising segments.

http://webpolicy.org/2014/10/24/how-verizons-advertising-header-works

Privacy, Anonymity et al

Some terminology:

- secrecy is the mechanism used to limit the number of principals with access to information (e.g., 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 (e.g., 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, exams)

Privacy vs Anonymity

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

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 Problems

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 Problems

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 Siri dictation (sent over the Internet to Apple; retained for 2 years)
- markets often only work, if data is restricted (to build trust)
- social networks can reveal data about you
- have you tried the collusion (lightbeam?) extension for FireFox?
- I do use Dropbox, store cards

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 search)

- 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 search)

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

"The reality":

- London Health Programmes lost in 2011 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 search)

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

"The reality":

- also in 2011, 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

- 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 search)

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

"The reality":

- in 2007, Gordon Brown needed to apologise for the loss of tax data of 25M people (a junior civil servant sent a CD in the mail, which got lost)

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
  - 15 Million active users in the UK; 500M tweets per day
  - 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 Attacks

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 Attacks

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)

Re-identification Attacks

- in 1990 medical databases were routinely made public with names removed, but birth dates, gender, ZIP-code were retained
- could be cross referenced with public voter registration data in order to find out what the medical record of the governor of Massachusetts was (in 1997 Latanya Sweeney)

- 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

AOL Search Queries

- In 2006, AOL published 20 million Web search queries collected of 650,000 users (names had been deleted)
- ...within days an old lady, Thelma Arnold, from Lilburn, Georgia, (11,596 inhabitants) was identified as user No. 4417749
- some of the queries that identified her away:
  - landscapers in Lilburn, Ga
  - 60 single men
  - nicotine effects on the body
  - ...

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)

We cannot exclude all Harm

- Analysis of a given data set teaches us that smoking causes cancer. Mary, a smoker, is harmed by this analysis: her insurance premiums rise. Mary's premiums rise whether or not her data are in the data set. In other words, Mary is harmed by the finding smoking causes cancer.

- ...of course she is also helped; she might quit smoking

We cannot exclude all Harm

Supervising queries will also not work in general:

- denying a request can already disclose information
- in general it is not decidable, whether a sequence of queries can identify a person

Differential Privacy

- Goal: Nothing about an individual should be learnable from the database that cannot be learned without access to the database.
- Differential privacy is a "protocol" which you run on some dataset X producing some output $O(X)$.
- You want to achieve forward privacy

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 $\quad$ tell me $f(x) \Rightarrow$ $\quad$ Database
$\Leftarrow f(x) + \text{noise}$ $\quad$ $x_1, \ldots, x_n$

- $f(x)$ 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

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

Example

Name	Has the disease?
Alice	yes
Bob	no
Charlie	yes
Eve	no
Chandler	yes

How many people have a disease?

Adding Noise

Adding noise is not as trivial as one would wish:

- If I ask how many of three have a disease 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 the disease, is male, below 30, ...), then one individual can change the dataset by 5

Differential Privacy Problems

- How to do differential privacy "offline" is still an active research question?
- What constitutes a single entry in the database?
- Evolution of a database:

Name	Has the disease?	
Alice	yes	
Bob	no	
Charlie	yes	
Eve	no	
Chandler	yes	
Marc	yes	$\Leftarrow$ new entry

Tor

??

Take Home Point

According to Ross Anderson:

- Creating large databases of sensitive personal information is intrinsically hazardous (NHS)
- Privacy in a big hospital is just about doable.
- How do you enforce privacy in something as big as Google or complex as Facebook? Nobody knows.

Similarly, big databases imposed by government.