Access Control and Privacy Policies (9)

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

Last Week

Recall, the Schroeder-Needham (1978) protocol is vulnerable to replay attacks.

$$
\begin{array}{l}
A \to S : A, B, N_A \\
S \to A : \{N_A, B, K_{AB}, \{K_{AB}, A\}_{K_{BS}}\}_{K_{AS}} \\
A \to B : \{K_{AB}, A\}_{K_{BS}} \\
B \to A : \{N_B\}_{K_{AB}} \\
A \to B : \{N_B - 1\}_{K_{AB}}
\end{array}
$$

Last Week

Recall, the Schroeder-Needham (1978) protocol is vulnerable to replay attacks.

$$
\begin{aligned}
A &\rightarrow S : A, B, N_A \\
S &\rightarrow A : \{N_A, B, K_{AB}, \{K_{AB}, A\}_{K_{BS}}\}_{K_{AS}} \\
A &\rightarrow B : \{K_{AB}, A\}_{K_{BS}} \\
B &\rightarrow A : \{N_B\}_{K_{AB}} \\
A &\rightarrow B : \{N_B - 1\}_{K_{AB}}
\end{aligned}
$$

Fix: Replace messages 2 and 3 to include a timestamp:

$$
\begin{aligned}
S &\rightarrow A : \{B, K_{AB}, T_S, \{K_{AB}, A, T_S\}_{K_{BS}}\}_{K_{AS}} \\
A &\rightarrow B : \{K_{AB}, A, T_S\}_{K_{BS}}
\end{aligned}
$$

Denning-Sacco Fix

Denning-Sacco (1981) suggested to add the timestamp, but omit the handshake:

$$\begin{array}{l} A \rightarrow S : A, B \\ S \rightarrow A : \{B, K_{AB}, T_S, \{K_{AB}, A, T_S\}_{K_{BS}}\}_{K_{AS}} \\ A \rightarrow B : \{K_{AB}, A, T_S\}_{K_{BS}} \\ B \rightarrow A : \{N_B\}_{K_{AB}} \\ A \rightarrow B : \{N_B - 1\}_{K_{AB}} \end{array}$$

they argue A and B can check that the messages are not replays of earlier runs, by checking the time difference with when the protocol is last used

Denning-Sacco-Lowe Fix of Fix

Lowe (1997) disagreed and said the handshake should be kept, otherwise:

$$\begin{array}{rl} A \rightarrow S : & A, B \\ S \rightarrow A : & \{B, K_{AB}, T_S, \{K_{AB}, A, T_S\}_{K_{BS}}\}_{K_{AS}} \\ A \rightarrow B : & \{K_{AB}, A, T_S\}_{K_{BS}} \\ I(A) \rightarrow B : & \{K_{AB}, A, T_S\}_{K_{BS}} \quad \text{replay} \end{array}$$

When is this a problem?

Denning-Sacco-Lowe Fix of Fix

Lowe (1997) disagreed and said the handshake should be kept, otherwise:

$$\begin{aligned}
A &\rightarrow S : A, B\\
S &\rightarrow A : \{B, K_{AB}, T_S, \{K_{AB}, A, T_S\}_{K_{BS}}\}_{K_{AS}}\\
A &\rightarrow B : \{K_{AB}, A, T_S\}_{K_{BS}}\\
I(A) &\rightarrow B : \{K_{AB}, A, T_S\}_{K_{BS}} \quad \text{replay}
\end{aligned}$$

When is this a problem?

Assume B is a bank and the message is “Draw £1000 from A's account and transfer it to I.”

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 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 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?

- Assume you make a survey of 100 randomly chosen people.
- Say 99% of the surveyed people in the 10 - 40 age group have seen the Gangnam video on youtube.

- What can you infer about the rest of the population?

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 $f(x) \Rightarrow$	Database
	$\Leftarrow f(x) +$ noise	$x_1, \ldots, x_n$

- $f(x)$ 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 $f(x) \Rightarrow$ $\Leftarrow f(x) +$ 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:
  Christian's salary $=$

  $$\Sigma \text{ all staff} - \Sigma \text{ all staff} \setminus \text{Christian}$$

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 $\mathbf{1}$. So answers would be in the range of $\mathbf{1}$ to $\mathbf{3}$

- 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 $\mathbf{5}$

Tor, Anonymous Webbrowsing

- initially developed by US Navy Labs, but then opened up to the world
- network of proxy nodes