What is Privacy?

- *The moral claim of individuals to be left alone and to control the flow of information about themselves* (Laudon 1996)
Privacy Concerns

- Tracking web browsing, search, messaging, etc.
- Health records (jobs, insurance, …)
- Photographs, video, etc.

Behavioral anecdote

More than 70% of people would reveal their computer password in exchange for a bar of chocolate, a survey has found.

It also showed that 34% of respondents volunteered their password when asked without even needing to be bribed.

A second survey found that 79% of people unwittingly gave away information that could be used to steal their identity when questioned.

Security firms predict that the lax security practices will fuel a British boom in online identity theft.

Security shock

The survey on passwords was carried out for the Infosecurity Europe trade show due to take place at Olympia in London from 27-29 April.
Behavioral anecdote

The survey data was gathered by questioning commuters passing through Liverpool Street station in London and found that many were happy to share login and password information with those carrying out the research.

As well as people simply telling the questioners their passwords or saying they would hand them over in exchange for some confectionery, a further 34% revealed the word or phrase they used when asked if it had anything to do with a pet or child’s name.

Family names, pets and football teams were all used by those questioned to provide inspiration for a password.

The survey found that, on average, people have to remember four passwords, though one unlucky respondent had to remember 40.

“Ability to control flow of information about oneself”
Economic context

• A data reseller.
• Acquires data about individuals:
  – Smart phone (conversations, location, browsing, etc.)
  – Medical records
  – Demographic
  – Family
  – Occupation
  – …
• Valuable for public health; BUT privacy concern.

Markets trading data about you

• BlueKai (now Oracle): a cloud-based big data platform that enables companies to personalize online, offline, and mobile marketing campaigns
• Axciom, eXelate, Datalogix, etc.
Markets trading data about you

• BlueKai (now Oracle): a cloud-based big data platform that enables companies to personalize online, offline, and mobile marketing campaigns
• Axciom, eXelate, Datalogix, etc.

• Today: focus not on targeting, but on statistical analysis of data.

Can we release data*, for example about recreational drug use, such that no information is revealed** about any single individual?

* While allowing useful statistical analysis
** With high probability
Privatized Data Release

• First idea: **Hide identifiers**
  – Hide (name, address, SSN)

Re-identification: Gov. Weld

• MA Group Insurance Commission, who purchases insurance for state employees, 1997 health data release
• William Weld (Governor) assured the public about privacy— they had removed identifiers.
• Left (ZIP, DoB, gender.) “quasi-identifier”
• Latanya Sweeney purchased Cambridge voter data for $20 (name, address, ZIP, DoB, gender).
• Has Weld’s ZIP, DoB, gender.
• NOW: this QI is UNIQUE in the GIC release. (Sweeney sent Weld’s health records to his office)
Re-identification: PGP

- Personal Genome Project (PGP): 2014
  - Sequence genotypic and phenotypic information of 100,000 informed volunteers, make data public

- PGP profile includes:
  - Medications, diagnoses, procedures, DNA
  - For 579 people, included ZIP, DoB and gender (from public records in PGP)

- Sweeney again used voter data and other public records to de-identify ~100 records.

Privatized Data Release

- First idea: Hide identifiers
  - Name, address, SSN

- Second idea: Hide indirect identifiers
  - gender, DoB, place of birth, ZIP code, ..
Re-identification: AOL search

• AOL released 20 million search queries on August 4, 2006.

A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.  AUG. 9, 2006

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher’s anonymity, but it was not much of a shield.

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from “numb fingers” to “60 single men” to “dog that urinates on everything.”

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for “landscapers in Lilburn, Ga.,” several people with the last name Arnold and “homes sold in shadow lake subdivision gwinnett county georgia.”

It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga., frequently researches her friends’ medical ailments and loves her three dogs.

“Those are my searches,” she said, after a reporter read part of the list to her.
Privatized Data Release

• First idea: Hide identifiers
  – Name, address, SSN
• Second idea: Hide indirect identifiers
  – gender, DoB, place of birth, ZIP code, ..

What is an “indirect identifier”
What data can be linked to re-identify someone?

Re-identification: Netflix Movies

• Netflix Challenge, 2006
• Data set: 100,480,507 ratings by 480,189 users on 17,770 movies (userID, movie, date, #stars)
• Goal: Improve recommender system accuracy by 10% ($1,000,000 prize). Won in 2009.
• A “linkage attack” was conducted in 2007, matching records with IMDb records (name, movie, date, review).
• Re-identified two users.
• Netflix settled a class action lawsuit in 2010
Privatized Data Release

• First idea: Hide identifiers
  – Name, address, SSN

• Second idea: Hide indirect identifiers
  – gender, DoB, place of birth, ZIP code, ...

• Third idea: \textit{k-anonymity} \hspace{1em} (Sweeney ’02)
  – Quasi-identifier = (gender, DoB, ZIP)
  – Make sure each QI appears $\geq k$ or zero times
  – Coarsen data as needed (e.g., bucket age)

\textit{k-anonymity} is insufficient

(1) Need to define indirect identifiers

(2) What if every user with the same QI may use recreational drugs?
k-anonymity is insufficient

(1) Need to define indirect identifiers

(2) What if every user with the same QI may use recreational drugs? (Insisting on diversity of sensitive values for records with same QI doesn’t help; what if attacker already knows status of other 7 people?)

What went wrong?

• Grade 11 test results reported by a high school in California.
• Redacted data when <10 people.

<table>
<thead>
<tr>
<th></th>
<th>M test took</th>
<th>M test passed</th>
<th>ELA test took</th>
<th>ELA test passed</th>
</tr>
</thead>
<tbody>
<tr>
<td>total</td>
<td>27</td>
<td>3</td>
<td>23</td>
<td>3</td>
</tr>
<tr>
<td>gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>4</td>
<td>*</td>
<td>3</td>
<td>*</td>
</tr>
<tr>
<td>M</td>
<td>23</td>
<td>3</td>
<td>20</td>
<td>2</td>
</tr>
<tr>
<td>race</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic or Latino</td>
<td>20</td>
<td>3</td>
<td>18</td>
<td>3</td>
</tr>
<tr>
<td>White</td>
<td>7</td>
<td>*</td>
<td>5</td>
<td>*</td>
</tr>
</tbody>
</table>
Privatized Data Release

• First idea: Hide identifiers
  – Name, address, SSN

• Second idea: Hide indirect identifiers
  – gender, DoB, place of birth, ZIP code, ..

• Third idea: $k$-anonymity
  – Quasi-identifier = (gender, DoB, ZIP)
  – Make sure each QI appears $\geq k$ or zero times
  – Coarsen data as needed (e.g., bucket age)

• Fourth idea: Differential Privacy

“Ability to control flow of information about oneself”

We need a good definition

What about:
“The knowledge of a data user about any individual is left almost unchanged by access to a database.”
• “The knowledge of a data user about any individual is left almost unchanged by access to a database.”

• Too strong!
  – Aliens come to earth…..

Differential privacy (Dwork et al. 06)

• **Differential privacy** (informal)

*Participation by an individual in a database leaves the knowledge of a data user about the individual almost unchanged*
Differential privacy  

(Dwork et al. 06)

• **Differential privacy** (informal)

  Participation by an individual in a database leaves the knowledge of a data user about the individual almost unchanged

• Define for query A(D) on a database
  – “How many people watch Jaws?”
  – “Give me the database.”

Differential privacy  

(Dwork et al. 06)

• D ~ D’ are “adjacent” if differ in one record
• *randomized query response* A(D)
• A is eps-DP, for small eps>0, if, for all D~D’, all sets of possible answers S:
Differential privacy (Dwork et al. 06)

• D ~ D’ are “adjacent” if differ in one record
• randomized query response $A(D)$
• $A$ is $\epsilon$-DP, for small $\epsilon > 0$, if, for all $D \sim D'$, all sets of possible answers $S$:

$$1 - \epsilon \leq \frac{\Pr(A(D) \in S)}{\Pr(A(D') \in S)} \leq 1 + \epsilon$$

E.g., how many people watch Jaws? Prob = 7 is similar on D and D’

Useful properties of Diff Priv.

• Belief about database D given answer $A(D)$ doesn’t depend much on D vs D’
• Probability of any downstream action doesn’t change by much. (e.g., robocall at dinner)
• Composition: query $A_1$ followed by query $A_2$ is $(\epsilon_1 + \epsilon_2)$-DP
• Any post-processing is ok!
How to achieve diff. privacy?

An idea: Noisy perturbation

• Add *zero-mean noise* to a query.
  – E.g., #people in a ZIP code who have watched Jaws

• Query response can be vector valued
  – E.g., give me a count of pass/fail test by gender and ethnicity

• Idea: any particular response is almost as likely for any pair of adjacent databases.
Query Sensitivity

• Amount of noise to add depends on query sensitivity (and desired eps-DP target)

\[ \Delta q = \max_{D', D'' \text{ s.t. } D' \sim D''} \| q(D') - q(D'') \|_1 \]

Query Sensitivity

• What is the sensitivity of the following:
  – \( q(D) \) is #people who have watched Jaws
  – \( q(D) \) is #times Jaws has been watched
  – \( q(D) \) is #people watched Jaws + #people who have watched ET + #people who have watched Star Wars
  – \( q(D) \) is #people who have watched two of Jaws, ET and Star Wars
  – \( q(D) \) is number 4* and 5* reviews of Jaws
Laplace Mechanism

Given query q, get Alg: \[ A^L_j(D) = q_j(D) + Z_j \]
\[ Z_j \sim \text{Lap} \left( \frac{\Delta q}{\epsilon} \right) \]

**Theorem:** \( A^L(D) \) is \( \epsilon \)-differentially private.

Idea of Laplace noise

- **Query:** \#users watching *Jaws*
- Perturb output by adding Laplace noise
- Consider \( D \) and \( D' \), \( D \sim D' \)

![Graph showing probability density](image-url)
Accuracy of Lap Mechanism

• Depends on sensitivity \(\Delta q\), epsilon, and dimension \(k\) of data.

• Say a query is “alpha-accurate w.p. 1-beta” if
  \[ \Pr \left( |q_j(D) - A^L_j(D)| < \alpha \right) \geq 1 - \beta \]

• Not too large in any one dimension of query

• **Thm.** Laplace mechanism is
  \[ (\Delta q/\epsilon) \ln(k/\beta)-accurate \quad \text{w.p. 1-beta.} \]
Interactive vs Non-interactive release

• Interactive: Need to know total #queries, and allocate epsilon “privacy budget”; e.g., eps/m per query
• Non-interactive: query is “give me the database”. Any number of queries on that database are then possible.

Non-interactive release

• A DB = a count of records of different types!

• Query: “partition the data into $k$ record types and count the number in each”
• E.g., number of people watching Jaws in each DoB, ZIP, gender type.
• Sensitivity = 1

• Granularity vs Accuracy tradeoff. Why?
Non-interactive release

• Query: “partition the data into $k$ record types and count the number in each”
• Example (eps=ln(2)/2)

\[
\begin{array}{c|c|c}
\text{Smoking}=&\text{N} & \text{Smoking}=&\text{Y} \\
\text{Cancer?}=\text{N} & 250 + \text{Lap}(\frac{2}{\ln(2)}) & 50 + \text{Lap}(\frac{2}{\ln(2)}) \\
\text{Cancer?}=\text{Y} & 50 + \text{Lap}(\frac{2}{\ln(2)}) & 250 + \text{Lap}(\frac{2}{\ln(2)})
\end{array}
\]

Can post-process to clean-up!

Diff. privacy is an active field

• For learning algorithms
  – clustering, regression, deep learning, …
  – regularization <> low sensitivity

• Connections to EconCS:
  – Mechanism design via diff. privacy
  – Modeling the utility of privacy, allowing tradeoffs of $$s$$s for privacy. Market design.
Summary

• Differential privacy looks like the right definition; has created an entire field…
  – Big effort at Harvard with Salil Vadhan, Cynthia Dwork, CRCS.
  – Privacy tools project (event on Mon Dec 11)
  – Interest in finding the right tradeoff between privacy and data utility

• Note: differential privacy doesn’t say anything about personalization/targeting.