Anonymity and databases

Privacy protection

- If databases containing private information are released, certain fields can be sanitized
  - SSNs, names, addresses, etc.
  - Certain types of collections of data, in combination, can also reveal identity: zip code, birth date, and gender, for example
- So inference attacks are still possible
  - Goal is to use database information as well as public information to learn more about underlying data
  - Or use combinations of the less sensitive data to infer identity and confidential information

Two different notions of privacy

- Identity disclosure occurs when an individual is linked to a particular record in the released table.
- Attribute disclosure occurs when new information about some individuals is revealed, i.e., the released data makes it possible to infer the characteristics of an individual more accurately than it would be possible before the data release.
  - Note that this often leads to identity disclosure, but not always.
- Either of these can cause harm.

Two examples

- Netflix and IMDB both released databases of user habits
    - Netflix left off all user info, but in IMDB that information can be left on (at user’s discretion)
    - Researchers at UT Austin managed to determine a Netflix user based on the IMDB data
- Medical encounter database:
    - Anonymized insurance database kept birthday, sex and zip code
    - Researcher from CMU linked this with voter registration records and found the medical record of the governor of Massachusetts

Protecting against inference attacks

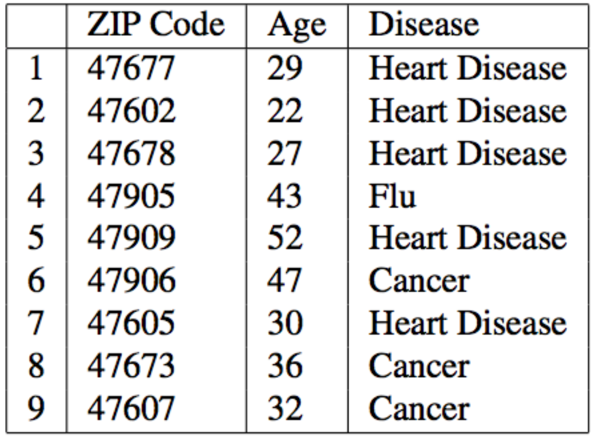
- Standard techniques:
  - Cell suppression: some cells are removed in the published version, to make inference attacks harder
  - Generalization: Instead of specific values, ranges are included in the released database
    - Example: Age range rather than specific age
  - Noise addition: Every value in the database has a small(ish) random number added to it
    - Goal: Average doesn’t change, but each individual entry does

Downside of obfuscation

- For each of these, the data becomes less specific, so there is a trade-off.
    - In the extreme, data is so blurred that it is useless.
- No widely accepted standard, and this is a hot area of research
- Many are focused on what formal requirements we should have, as well as more and more sophisticated attacks.

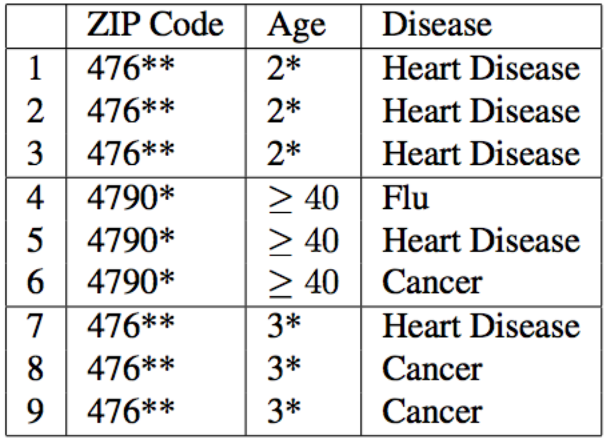
K-anonymization

- Database is secure if any possible SELECT query will return at least k records, where k is some threshold
- Often accomplished by adding fake data to the database
    - But as little fake data as possible
- One of the earliest notions of how privacy is "good enough" in a database
    - Heavily criticized but still used
- For example: If Alice knows that Bob is a 27-year old man living in ZIP 47678 and Bob's record is in the table. From Table 2, Alice can conclude that Bob corresponds to one of the first three records, and thus must have heart disease. [Li et al, 2007]



	ZIP Code	Age	Disease
1	47677	29	Heart Disease
2	47602	22	Heart Disease
3	47678	27	Heart Disease
4	47905	43	Flu
5	47909	52	Heart Disease
6	47906	47	Cancer
7	47605	30	Heart Disease
8	47673	36	Cancer
9	47607	32	Cancer

Table 1. Original Patients Table



	ZIP Code	Age	Disease
1	476**	2*	Heart Disease
2	476**	2*	Heart Disease
3	476**	2*	Heart Disease
4	4790*	≥ 40	Flu
5	4790*	≥ 40	Heart Disease
6	4790*	≥ 40	Cancer
7	476**	3*	Heart Disease
8	476**	3*	Cancer
9	476**	3*	Cancer

Table 2. A 3-Anonymous Version of Table 1

Li, Li, and Venkatasubramanian 2007

L-diversity

- To address limitations, l-diversity was introduced:
  - An equivalence class of entries is l-diverse if there are at least l “well-represented” values for any sensitive attribute.
  - By well-represented, we could perhaps mean there are at least l distinct values for the sensitive attributes in each equivalence class.
  - (Note that there are other, more complex definitions, but all attempt to minimize how you isolate entries.)
- However, still not enough in all cases:
  - In prior example: Suppose that one equivalence class has an equal number of positive records and negative records. This STILL presents a serious privacy risk, because anyone in the class would be considered to have 50% possibility of being positive, as compared with the 1% of the overall population.

Differential privacy

- For any record R in the database and sensitive property P, the probability p that R is in the database and the probability p’ that R is not in the database differ by at most some ε
  - Essentially, given two very similar databases, the probability that a query will look the same from each is very high
- Considerably more sophisticated, but also harder to work with.

Taking a step back

- None of this even touches on how companies use internal, collected data in unexpected ways.
- Target, for example: “Target is renowned in the retail world for its data collection and analysis, grabbing bits and pieces wherever it can - from your store purchases to visits to its website to surveys you’ve taken to things you’ve posted on Facebook.”
  - They have entire teams dedicated to matching your personal info to your shopping habits, so they can better advertise.
  - “Andew Pole, who heads a 60-person team at Target that studies customer behavior, boasted at a conference in 2010 about a proprietary program that could identify women - based on their purchases and demographic profile - who were pregnant.”

Source: “What Target Knows About You”, Reuters 2014

Opting out?

- In 2014, journalist Janet Vertesi attempted to “opt out”:
  - She contacted family and friends and asked them not to post or message anything on social media about the pregnancy.
  - She downloaded Tor, and shopped only with it.
  - She purchased all baby related things with cash, and even sent up an Amazon account that linked to a personally hosted email, purchased gift cards with cash, and had them delivered to a PO Box.
- The result:
  - “For months I had joked to my family that I was probably on a watch list for my excessive use of Tor and cash withdrawals. But then my husband headed to our local corner store to buy enough gift cards to afford a stroller listed on Amazon. There, a warning sign behind the cashier informed him that the store “reserves the right to limit the daily amount of prepaid card purchases and has an obligation to report excessive transactions to the authorities.””

Today’s in class exercise

- Think about how social media collects and shares some aspects of your account. They don’t release private info, but (for example) might collect information about what you’ve liked and linked to.
- First: identify 3 aspects of your social media behavior that when linked together, would give a higher probability of identifying you.
- Then: reflect on how personalized marketing has affected you. What ads or coupons do you get? Do you think these algorithms are effective?
- Finally: Do you find the marketing “creepy”, or useful?