Privacy in a Networked World: Trouble with Anonymization, Aggregates

Historical US Privacy Laws

- First US Law dates back to: 1890
- Protecting privacy of Individuals against government agents
  - 1973 report.
- Restricting access to personal information
  - Family Education Rights and Privacy Act
  - Right to Financial Privacy Act
  - Health Insurance Portability and Accountability Act
- Curtailing intrusion into homes
  - Third Amendment: "No quarter"
  - Fourth Amendment: "No search/seizure"

Privacy in a Networked World

What happens when internet companies store browsing information about you?

Not the government.

Unclear if this is sensitive.

Unclear if they invaded your home.

AFLB ~ Early 2006

- AFLB: Algorithms for the lunch bunch
- Distinguished speaker from Google
- Excitement among Stanford students---promise of practically relevant problems

Exasperation

Stanford Prof: "How about approach X?"

Senior Googler: "Your proposal won't work. We've already tried it."

Stanford Prof: "What else have you tried?"

Senior Googler: "We cannot tell you."

Stanford Prof. "Can you give us data? So we can try out our ideas?"

Senior Googler: "Impossible."

Stanford Prof: "So you have told us about some interesting problems, but there is no way we can make progress on any of them. Why are you here?"

Moral of the Story

- Data needed to evaluate ideas
- Even anonymous data would be okay
- Researchers wanted to improve search algorithms not delve into users' lives
- Not the complete story as we shall see

Summer in Seattle

- I interned at Microsoft Research, Silicon Valley in the Summer of 2007
- Visited AdCenter Labs a few times over the summer
- No rain, lots of data => happiness

"Don't get me fired"

- The sign on Ying Li's door with a printout of NY Times article
- Fittingly, Ying Li heads a security and privacy division today.
- Queries in the NYT article are benign, but can be disturbing

An Attack on Health Information

- Group Insurance Commission in MA responsible for buying health insurance of state employees

- They collected gender, zip-code, ethnicity, age, diagnosis, procedure, medication for employees
- They believed this information is anonymous, so they sold copy to industry; Notice this information is useful for various reasons

Sweeney's Observation

87% of the population in the United States had reported characteristics that likely made them unique based only on ZIP, gender, date of birth

Try this out: http://aboutmyinfo.org/index.html

Next..

- Sweeney buys voter registration record for 20$
- Upshot: Public medical records!

Moral of the Story

- Removing username, ssn does not anonymize data

- Cannot release data set as is
- Search Logs Impossibility Equation:

Happy users + researchers = impossible

unless researcher is at search company :P

Some Successes

- Google Trends, Facebook data science
  - trends of individual query terms/behavior of many people

- Specific application: Flu Trends
  - early detection of disease activity, when follow
    a rapid response, can reduce the impact of bo
    seasonal and pandemic influenza

k-anonymity

- Prevents 'linkage' attacks of the kind discussed so far
- Definition Parsing Challenge :)
  - Given a population of entities U, an entity-specific table $T(A_1 \ldots A_n)$, $f: U \rightarrow T$ and $g: T \rightarrow U'$, where $U \subseteq U'$. A quasi-identifier of T, written Q_T, is a set of attributes $\{A_i \ldots A_j\} \subseteq \{A_1 \ldots A_n\}$ where: $\exists\, p_i \in U$ such that $g(f(p_i)[Q_T]) = p_i$

Definitions Reloaded

A *quasi-identifier* is a set of attributes that uniquely identify at least one individual in a public database.

Let A' be the union of the set of quasi-identifiers.

A 'private' table satisfies k-anonymity if every combination of values of the attributes in A' that occurs in table, occurs at least k times.

Examples

2-anonymous table

Is this table 3-anonymous?

Methods of Achieving k-anonymity

- Generalization
  - 94085 --> 94***
- Suppression
  - Japanese --> *

See Figures 1,2 on page 2 of this paper.

Some Issues (Machanavajjhala 2007)

Attack #1:

Homogeneity attack: All k people have AIDS

Attack #2:

Suppose k =4. In one group, two people have AIDS, two have hypertension. You have a Japanese friend who falls in that group. Suppose Japanese have low incidence of hypertension...

Lessons so far

Privacy is hard to get right!

Next lecture, a more airtight definition of privacy, browser privacy social network privacy..

End of Lecture 1