Data Aggregation, Privacy Threats and Anonymity

WSPC, chapter 8

Privacy concerns

- The technologies of the Internet and the Web was designed to **transfer information**, **not protect the privacy** of people who use this information;
- Moreover, existing technologies provide with opportunities to collect a lot of information about users of computer networks
- Example: a web site may collect information about people seeing their messages:
  - Where they live;
  - What other web sites the person has visited;
  - Their email addresses;
  - etc

Information privacy

- Alan Westin (1967), **Information privacy**:
  "the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others."
- **Information privacy threat**: many individuals and institutions have lost ability to control how and to what extent information about them is communicated to marketing companies, government agencies, etc

Personal and Private Information

- **Personal information**. Information about a person, like a name, date of birth, names of parents, attended schools, etc
- **Private information.** Personal information that is not generally known. Some kinds of private information, for example bank records are protected by law.

Personally identifiable and anonymized information

- **Personally identifiable information.** Information from which a person identity can be derived. Example: an account number.
- **Anonymized information.** Information from which a person identity cannot be derived. Example: an age of person (if no other information is available)

Aggregate information

- Statistical information combined from many individuals to form a single record. **Example:** national Census Bureau.
  - No person identity can be extracted from aggregate information alone, but when combined with other anonymized information, aggregate information can help to identify and reveal particular characteristics of an individual

Privacy threats based on anonymized information

Combining seemingly anonymous information one may reveal identity with high probability:

If you ask a person for

- a birthday
- Zip code (in US)
- an age

you actually ask for personally identifiable information, even though it looks as anonymized.

S.Garfinkel in WSPC: an average 8 people in each zip code have the same birthday.

"Personal data for the Taking"

- The title above is the title of the article in the *New Your Times*, by Tom Zeller, Jr, published May 18, 2005
- It analyses the project at John Hopkins University, done by Prof. Aviel D. Rubin and his students
- Several groups of three to four students set out to collect personal information on citizens of Baltimore using only legal, public sources of information, such as death records, property tax, campaign donations, occupational registries, etc
- Each group could spend no more that $50
- Altogether they gathered over a million records with hundreds of thousands of individuals

"Personal data for the Taking"(cont)

- Databases they collected were cleaned an linked, making it possible to query a multiple layers of information about a single name;
- Typical result about an individual:
  - Name
  - Precise address
  - Occupation (with the details of his professional license)
  - The name of his wife
  - Their birth dates
  - The price of the home they paid
  - The party registration
  - Elections he has voted in last 25 years

"Personal data for the taking"(cont.)

Lessons:

- A lot of personal information easily available on the Web;
- When consolidated it may violate privacy of individuals and be used e.g. for stealing identity
- Competing social interests in openness and privacy
- No ready solution

Sources of personal and private information

- User-provided information
  - When using online services, or buying online users provide personal information, such like names, addresses, passwords, additional passwords, sometime date of birth, etc
- Information obtained by observations of users activities, or traces they left
  - Log Files: Web logs, Mail Logs, DNS logs etc
  - Cookies
  - Web Bugs, Adware, Spyware

Anonymity in communications

- One of the ways to protect privacy is to make a communication **anonymous,** so an adversary that
  - can monitor and/or compromise certain parts of the systems
  - would not be able to match a message (request) sender with the recipient (sender-recipient matching).
- Most widespread methods for anonymity in the Web communications based on the idea of **third trusted party** serving as **anonymizer** (special proxy server)

Anonymizer

- Examples: Anonymizer.com
- Upon a request from an user anonymizer fetches the web page and displays is within your browser
- In doing so, anonymizer does not leave information about your request on the web-server: it re-directs your request, replacing all sensitive information (IP address, etc) with its own details
- Additionally it may provide
  - encryption of traffic between an user and itself
  - Blocking and removing potential active privacy and security threats: web bugs, spyware, viruses, etc

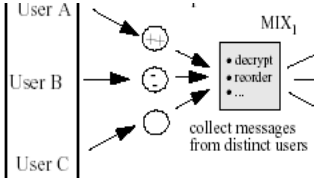
Anonymizer

- Good protection and reasonable cost
- Privacy protection is based on the trusted central proxy
- Central proxy "knows everything" about communication – attacks by "insiders" are possible

Mix-Networks

D.Chaum (1981):

A mix node is a node in the network that takes a number of incoming messages (packets), modifies them and output in a random order



Mix-networks

The mix nodes can be used for anonymous communication as follows:

- The message will be sent through a sequence of mix nodes (a route) p_1, p_2, p_3, …p_d. The user encrypts the message with node p_d key, the result encrypt with the node p_{d-1} key, etc
- Every mix node receives several messages, decrypt them, re-order and send to the next nodes in the route
- Every nodes "knows" only previous and the next node in the route => compromising a single, or even several (not all) mix nodes does not give an attacker an information about sender-receiver matching
- It is more expensive than anonymizer solution but gives more privacy protection. Example: *Freedom* at www.freedom.net