De-Identification and Re-Identification

Pierangela Samarati
Dipartimento di Informatica
Università degli Studi di Milano
pierangela.samarati@unimi.it

Algoritmi per l'Integrazione di Dati Sanitari:
Database, Record-linkage, Anonimizzazione

Milan, Italy – April 14, 2016

Motivation – 1

- Continuous growth of:
  - digital representation of information for more efficient and effective processing and sharing
  - need to cooperate and share data
  - government and company databases
  - user-generated content delivered through collaborative Internet services such as YouTube, Flickr
  - personally identifiable information collected whenever a user creates an account, submits an application, signs up for newsletters, participates in a survey, ...

Motivation – 2

- Data sharing and dissemination:
  - provide services
  - study trends or to make useful statistical inference
  - share knowledge
- External data storage and computation:
  - cost saving and service benefits
  - higher availability and more effective disaster protection
- Need to ensure data privacy is properly protected

Data sharing/publication – 1

- Statistical DBMSs: the DBMS responds only to statistical queries (e.g., avg, sum, count, ...)
- Statistical data (macrodata): release of pre-computed statistics (e.g., count/frequency or magnitude tables)
- Microdata: individual records are released

Data sharing/publication – 2

Need to guarantee confidentiality of sensitive information

- Identity disclosure: record in an anonymized dataset can be linked with a respondent's identity (problematic for microdata, less so for macro data)
- Attribute disclosure: the value of a confidential attribute of a respondent can be determined more accurately with access to the released dataset
- Inferential disclosure: information can be inferred with high confidence from statistical properties of the released data

Macrodata vs microdata

- In the past data were mainly released in tabular form (macrodata) and through statistical databases
- Today many situations require that microdata be released
  - increased flexibility and availability of information for the users
- Microdata are subject to a greater risk of privacy breaches (linking attacks)

Anonymization

- Datasets truly anonymized are not subject to privacy regulations
- Anonymization should ensure that nobody can:
  - single out an individual
  - link two records in a dataset
  - infer any information in such dataset

$\Longrightarrow$ removing directly identifying information is not enough

The anonymity problem

- The amount of privately owned records that describe each citizen's finances, interests, and demographics is increasing every day
- These data are de-identified before release, that is, any explicit identifier (e.g., SSN) is removed
- De-identification is not sufficient
- Most municipalities sell population registers that include the identities of individuals along with basic demographics
- These data can then be used for linking identities with de-identified information $\Longrightarrow$**re-identification**

The anonymity problem – Example

SSN	Name	Race	DoB	Sex	ZIP	Marital status	Disease
		asian	64/04/12	F	94142	divorced	hypertension
		asian	64/09/13	F	94141	divorced	obesity
		asian	64/04/15	F	94139	married	chest pain
		asian	63/03/13	M	94139	married	obesity
		asian	63/03/18	M	94139	married	short breath
		black	64/09/27	F	94138	single	short breath
		black	64/09/27	F	94139	single	obesity
		white	64/09/27	F	94139	single	chest pain
		white	64/09/27	F	94141	widow	short breath

Name	Address	City	ZIP	DOB	Sex	Status
.....	.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....	.....
Sue J. Doe	900 Market St.	San Francisco	94142	64/04/12	F	divorced
.....	.....	.....	.....	.....	.....	.....

Classification of attributes in a microdata table

- Identifiers: attributes that uniquely identify a data subject (e.g., SSN uniquely identifies the person with which it is associated)
- Quasi-identifiers: attributes that, in combination, can be linked with external information to reidentify all or some of the data subjects to whom information refers or reduce the uncertainty over their identities (e.g., DoB, ZIP, and Sex)
- Confidential: attributes that contain sensitive information (e.g., Disease)
- Non confidential: attributes that the data subjects do not consider sensitive and whose release does not cause disclosure

Re-identification

A study of the 2000 census data reported that the US population was uniquely identifiable by:

- year of birth, 5-digit ZIP code: 0.2%
- year of birth, county: 0.0%
- year and month of birth, 5-digit ZIP code: 4.2%
- year and month of birth, county: 0.2%
- year, month, and day of birth, 5-digit ZIP code: 63.3%
- year, month, and day of birth, county: 14.8%

Some microdata protection approaches

- k-anonymity: protects identity of respondents by confusing it in a set of at least k respondents
- ℓ-diversity: builds on k-anonymity adding condition that every computed group of respondents be associated with at least ℓ diverse occurrences of sensitive attributes
- t-closeness: builds on k-anonymity adding condition that distribution of sensitive attributes in every computed group of respondents be close to the one to be expected
- differential privacy: no respondent should make a difference on the result (adds noise to data)
- ...

k-Anonymity

Captures the following requirement:

- the released data should be indistinguishably related to no less than a certain number of respondents

Translating it to:

- each release of data must be such that every combination of values of quasi-identifiers can be indistinctly matched to at least k respondents
- publish only truthful information
- it guarantees so by generalizing (i.e., publishing at a lower level of detail) or suppressing quasi-identifying values

2-anonymized quasi-identifiers – Example

Race	DOB	Sex	ZIP
asian	64/04/12	F	94142
asian	64/09/13	F	94141
asian	64/04/15	F	94139
asian	63/03/13	M	94139
asian	63/03/18	M	94139
black	64/09/27	F	94138
black	64/09/27	F	94139
white	64/09/27	F	94139
white	64/09/27	F	94141

Race	DOB	Sex	ZIP
asian	64/04	F	941**
asian	64/04	F	941**
asian	63/03	M	941**
asian	63/03	M	941**
black	64/09	F	941**
black	64/09	F	941**
white	64/09	F	941**
white	64/09	F	941**

ℓ-Diversity

k-anonymity protects only identities, not the association with sensitive attributes, vulnerable to:

- homogeneity attacks: all respondents in a group have the same value for sensitive attributes
- background knowledge attacks: observers can rule out some possible associations based on other knowledge

Race	DOB	Sex	ZIP	Disease
...	...	...	...	...
black	64	F	941**	short breath
black	64	F	941**	short breath
...	...	...	...	...

ℓ-diversity: every group should contain at least ℓ well represented values

t-closeness

ℓ-diversity does not consider semantics and distribution of sensitive values

- skewness attacks: distribution of sensitive values in a group is different wrt the one of original population (e.g., 75% diabetes against 25%)
- similarity attacks: even if different, some values may be semantically similar (e.g., stomach ulcer and gastritis)

t-closeness: every group should have a distribution of sensitive values close to the distribution of the whole original population

Differential privacy

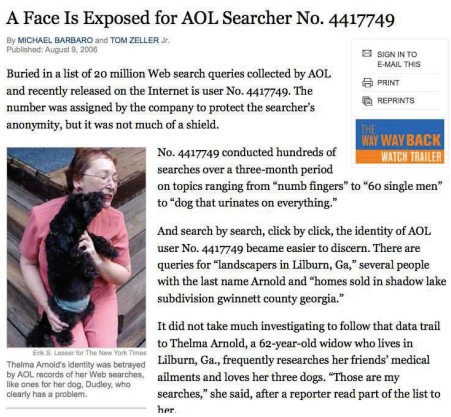
- Considers release/query of aggregate data
- Protects privacy by ensuring that no single person's inclusion or exclusion from the database can significantly affect the results of queries
- Provides privacy by adding noise
  - Difficult trade-off between privacy and utility

Anonymization is a complex problem ...

- Actions/logs can help re-identification
- Even pseudonyms can expose users
  - AOL: queries and pseudonyms
  - Netflix: linking to IMDb and pseudonyms
- Multiple sources
- Multiple releases

AOL data release – 1

- In 2006, AOL publicly posted to a website 20 million search queries for 650,000 users of AOL's search engine summarizing three months of activity
- AOL replaced identifying information (e.g., AOL username, IP address) with unique identification numbers (this made searches by the same user linkable)
- User 4417749:
  - "numb fingers", "60 single men", "dog that urinates on everything", "landscapers in Lilburn, Ga", "Arnold" (several people with this last name)



A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

SIGN IN TO E-MAIL THIS
PRINT
REPRINTS

THE WAY WAY BACK
WATCH TRAILER

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.

Erik S. Lesser for The New York Times
Thelma Arnold's identity was betrayed by AOL records of her Web searches, like ones for her dog, Dudley, who clearly has a problem.

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.

Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga

AOL data release – 2

What about user 17556639?

- how to kill your wife
- how to kill your wife
- wife killer
- how to kill a wife
- poop
- dead people
- pictures of dead people
- killed people
- dead pictures
- dead pictures
- dead pictures
- murder photo
- steak and cheese
- photo of death
- photo of death
- death
- dead people photos
- photo of dead people
- www.murderdpeople.com
- decapatated photos
- decapatated photos
- car crashes3
- car crashes3
- car crash photo

Netflix

- Only a sample of the movie ratings database was released
- Some ratings were perturbed (but not much to not alter statistics)
- Identifying information (e.g., username) was removed, but a unique user identifier was assigned
- De-identified Netflix data can be re-identified by linking with external sources (e.g., user ratings from IMDb users)

THREAT LEVEL | privacy

Netflix Spilled Your *Brokeback Mountain* Secret, Lawsuit Claims

BY RYAN SINGEL 12.17.09 4:29 PM

Follow @ryangel

Share 174 Tweet 18 +1 0 Share 5

An in-the-closet lesbian mother is suing Netflix for privacy invasion, alleging the movie rental company made it possible for her to be outed when it disclosed insufficiently anonymous information about nearly half-a-million customers as part of its $1 million contest to improve its recommendation system.

The suit known as Doe v. Netflix (.pdf) was filed in federal court in California on Thursday, alleging that Netflix violated fair-trade laws and a federal privacy law protecting video rental records, when it launched its popular contest in September 2006.

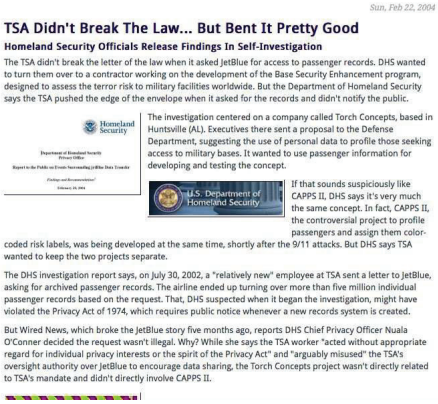
The suit seeks more than $2,500 in damages for each of more than 2 million Netflix customers.

In order to get a better movie recommendation algorithm, the online DVD rental company gave more than 50,000 Netflix Prize contestants two massive datasets. The first included 100 million movie ratings, along with the date of the rating, a unique ID number for the subscriber, and the movie info. Based on this data from 480,000 customers, contestants had to come up with a recommendation algorithm that could predict 10 percent better than Netflix how those same subscribers rated other movies.

Movies may reveal your political orientation, religious views, or sexual orientation

JetBlue

- In 2003, JetBlue Airways Corporation gave the travel records of five million customers to Torch Concepts (a private DoD contractor) for an antiterrorism study to track high-risk passengers or suspected terrorists
- Torch Concepts purchased additional customer demographic information (e.g., SSN) about these passengers from Axciom, one of the largest data aggregation companies in the U.S.
- The information from JetBlue and Axciom was then used by Torch Concepts to develop passenger profiles
- Claims of violation of JetBlue Privacy Policy



Sun, Feb 22, 2004

TSA Didn't Break The Law... But Bent It Pretty Good

Homeland Security Officials Release Findings In Self-Investigation

The TSA didn't break the letter of the law when it asked JetBlue for access to passenger records. DHS wanted to turn them over to a contractor working on the development of the Base Security Enhancement program, designed to assess the terror risk to military facilities worldwide. But the Department of Homeland Security says the TSA pushed the edge of the envelope when it asked for the records and didn't notify the public.

Homeland Security

Department of Homeland Security
Privacy Office
Report to the Public on Events Surrounding jetBlue Data Transfer
February 20, 2004

The investigation centered on a company called Torch Concepts, based in Huntsville (AL). Executives there sent a proposal to the Defense Department, suggesting the use of personal data to profile those seeking access to military bases. It wanted to use passenger information for developing and testing the concept.

U.S. Department of Homeland Security

If that sounds suspiciously like CAPPS II, DHS says it's very much the same concept. In fact, CAPPS II, the controversial project to profile passengers and assign them color-coded risk labels, was being developed at the same time, shortly after the 9/11 attacks. But DHS says TSA wanted to keep the two projects separate.

The DHS investigation report says, on July 30, 2002, a "relatively new" employee at TSA sent a letter to JetBlue, asking for archived passenger records. The airline ended up turning over more than five million individual passenger records based on the request. That, DHS suspected when it began the investigation, might have violated the Privacy Act of 1974, which requires public notice whenever a new records system is created.

But Wired News, which broke the JetBlue story five months ago, reports DHS Chief Privacy Officer Nuala O'Conner decided the request wasn't illegal. Why? While she says the TSA worker "acted without appropriate regard for individual privacy interests or the spirit of the Privacy Act" and "arguably misused" the TSA's oversight authority over JetBlue to encourage data sharing, the Torch Concepts project wasn't directly related to TSA's mandate and didn't directly involve CAPPS II.

Privacy and genomic data

The 1000 Genomes Project: international project (2008) to establish a catalogue of human genetic variation

- Five men involved in both the 1000 Genomes Project and a project that studied Mormon families from Utah were re-identified
  - their identities were determined
  - identities of their male and female relatives were also discovered
  - attack exploited haplotypes of short tandem repeats on the donor's Y chromosome

The Target Case

- Target assigns every customer a Guest ID number
- Analysts at Target identified 25 products that assign each shopper a pregnancy prediction score
  - woman, 23 y.o., buying in March cocoa-butter lotion, a purse large enough to double as a diaper bag, zinc and magnesium supplements and a bright blue rug
    ⇒ 87% due late August

Social networks

- People tend to connect with others with similar interests / activities / experiences . . .
- What one discloses exposes not only him/her but also others
  - a study in 2009 on 1,500 Facebook users showed that homosexual men have more homosexual friends than heterosexual men
  - tool to automatically predict the sexual orientation of Facebook users (not indicating it) based on their friends' orientations

Conclusions

- Sharing and easy access to information provides great benefits
- Need to ensure privacy of sensitive information is properly protected
- Problem is complex and needs to be handled with care