Data Anonymization Related Laws in the US and the EU

CS and Law Project Presentation

Jaspal Singh

The Need for Anonymization

- To share a database packed with sensitive information with third parties or with the entire public
  - Hospitals sharing health data with researchers
  - Websites selling transaction data to advertisers
- Need to **anonymize** the data to
  - protect the privacy of data subjects
  - ensure high utility

Naive Anonymization or the De-identification Process

- Remove personal identifiers like name, SSN, etc
- Many believe this process works perfectly in ensuring utility while maintaining privacy
- Many studies have shown it to be insufficient to protect privacy of data subjects
- Re-identification/Deanonymization - process of identifying an individual in anonymized data

Anonymized Databases Undone

- A study by Latanya Sweeney on 1990 census data discovered
  - (birth date, gender, zipcode) uniquely identify 87.1% individuals in the US
- Group Insurance Commission - a Massachusetts based govt. agency
  - purchased health insurance for state employees
  - decided to release records summarizing every state employee's hospital visits
  - removing fields containing name, address, social security number
  - William Weld, then Governor of Massachusetts, assured the public that patient privacy was protected
  - Dr. Sweeney purchased the complete voter rolls from the city of Cambridge

Anonymized Databases Undone

- The America Online (AOL) data release
- The Netflix Prize data study

Naive Anonymization or the De-identification Process

- Re-identification/Deanonymization - process of identifying an individual in anonymized data
- Re-identification techniques
  - Outside information
- More privacy preserving data anonymization techniques have been designed over time

Naive Anonymization or the De-identification Process

Name	Race	Birth Date	Sex	ZIP Code	Complaint
Sean	Black	9/20/1965	Male	02141	Short of breath
Daniel	Black	2/14/1965	Male	02141	Chest pain
Kate	Black	10/23/1965	Female	02138	Painful eye
Marion	Black	8/24/1965	Female	02138	Wheezing
Helen	Black	11/7/1964	Female	02138	Aching joints
Reese	Black	12/1/1964	Female	02138	Chest pain
Forest	White	10/23/1964	Male	02138	Short of breath
Hilary	White	3/15/1965	Female	02139	Hypertension
Philip	White	8/13/1964	Male	02139	Aching joints
Jamie	White	5/5/1964	Male	02139	Fever
Sean	White	2/13/1967	Male	02138	Vomiting
Adrien	White	3/21/1967	Male	02138	Back pain

Naive Anonymization or the De-identification Process

Race	Complaint
Black	Short of breath
Black	Chest pain
Black	Painful eye
Black	Wheezing
Black	Aching joints
Black	Chest pain
White	Short of breath
White	Hypertension
White	Aching joints
White	Fever
White	Vomiting
White	Back pain

Naive Anonymization or the De-identification Process

Race	Complaint
Black	Short of breath
Black	Chest pain
Black	Painful eye
Black	Wheezing
Black	Aching joints
Black	Chest pain
White	Short of breath
White	Hypertension
White	Aching joints
White	Fever
White	Vomiting
White	Back pain

Race	Birth Year	Sex	ZIP Code*	Complaint
Black	1965	Male	021*	Short of breath
Black	1965	Male	021*	Chest pain
Black	1965	Female	021*	Painful eye
Black	1965	Female	021*	Wheezing
Black	1964	Female	021*	Aching joints
Black	1964	Female	021*	Chest pain
White	1964	Male	021*	Short of breath
White	1965	Female	021*	Hypertension
White	1964	Male	021*	Aching joints
White	1964	Male	021*	Fever
White	1967	Male	021*	Vomiting
White	1967	Male	021*	Back pain

Anonymized related laws in the US

- No single set of data protection laws in the US
- Data protection laws are a combination of some federal and state law
- Different Acts in place to protect different types of data
- In general, the US data protection law assumes that the process of de-identification to maintain the privacy of the data subjects

Anonymized related laws in the US

Health Insurance Portability and Accountability Act (HIPAA)

- Introduced in 1996 with the aim to improve healthcare and health insurance in this country
- “de-identification of health information” (DHI)
- DHI - information that “does not identify an individual”
  - Suppress or generalize 18 identifiers
- HIPAA itself exempts data protected by DHI from any regulation

Anonymized related laws in the US

- Driver’s Privacy Protection Act
  - special protection for “personal information” including - SSN, driver identification number, name, address, telephone number
  - much less protection of information including - the 5-digit zip code, information on vehicular accidents, driving violations
- Federal Education Rights and Privacy Act (FERPA)
  - Enforces protection to directory info including name, address, telephone number, place of birth and major field of study
- Federal Drug Administration Regulations
  - Permits disclosure of “records about an individual” associated with clinical trials after deleting the names and other identifying information

Anonymized related laws in the EU

Article 2(a) of EU Data Protection Directive of 1995

- Personal data is “any information related to a natural person, who is identified or identifiable **directly or indirectly** in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity.”
- Anonymous data - complement of personal data
- The principles of protection shall not apply to data rendered anonymous

Anonymized related laws in the EU

Article 2(a) of EU Data Protection Directive of 1995

- To determine, whether a person is identifiable, account all means likely reasonably to be used either by the controller or by any other person to identify the said person
- Provision takes account of the fact that absolute anonymity is not achievable
- Anonymity is not static: the same information can be anonymous in one context and personal data in another

Anonymized related laws in the EU

Article 29 Working Party's Opinions on Anonymization (2014)

- Body of EU data protection regulators
- Helps interpret a legal criterion with applicable technical solutions
- Discusses on several anonymization techniques:
  - noise addition, substitution, aggregation, l-diversity, differential privacy

The new General Data Protection Regulation:

- Substantially similar definition of personal and anonymous data
- Concept of personal data made for specific and broadened

Discussion and Conclusion

- Absolute anonymity is not achievable
- Need to take into account:
  - Available technical solutions
  - Risks
  - Use cases of the released data
- US anonymization laws:
  - Aim at hiding personally identifiable information (PII) from the data
  - Different set of laws for different types of data
  - All laws assume the release-and-forget model of anonymization

Discussion and Conclusion

- EU anonymization laws:
  - Keeping pace with the advancements in the field of anonymization
  - Recommend a list of technical solutions
  - Does not distinguish between different types of data on the basis of usability
- A closer look at each type of data and application setting - check applicability of differential privacy
- No specific laws regarding anonymization of network data

Questions