

# Introduction to Privacy and Anonymity

MIE223  
Winter 2025

## 1 Privacy and Anonymity

### 1.1 AOL Privacy Debacle

- In August 2006, AOL released anonymized search query logs
  - 657K users, 20M queries over 3 months (March-May)
- Opposing goals
  - Analyze data for research purposes, provide better services for users and advertisers
  - Protect privacy of AOL users
    - \* Government laws and regulations
    - \* Search queries may reveal income, evaluations, intentions to acquire goods and services, etc.

### 1.2 AOL User 4417749

- AOL query logs have the form  $\langle \text{AnonID}, \text{Query}, \text{QueryTime}, \text{ItemRank}, \text{ClickURL} \rangle$ 
  - ClickURL is the truncated URL
- NY Times re-identified AnonID 4417749
  - Sample queries: “numb fingers”, “60 single men”, “dog that urinates on everything”, “landscapers in Lilburn, GA”, several people with the last name Arnold
    - \* Lilburn area has only 14 citizens with the last name Arnold
  - NYT contacts the 14 citizens, finds out AOL User 4417749 is 62-year-old Thelma Arnold

### 1.3 Foundations of Privacy

- Consent:
  - GDPR (EU), US (Privacy Act of 1974)
- Notice: you have to accept collection practices
  - Question: who are some of the major providers of user web data?
- De-identification
  - Only release attributes that could not identify you
  - Historically founded on principle of k-anonymity
    - \* Re-identification: multiple attributes can as well as quasi- identifiers (partial postal code) that link you accross datasets (medical, voter) even with k-anonymity
      - Sweeney (Harvard) used 1990 Census data to estimate that 0.04 percent of the United States population was uniquely identified by the basic demographic fields allowed by the HIPAA Safe Harbor – namely, year of birth, gender, and first 3 digits of ZIP

## 1.4 Background

- Large amount of person-specific data has been collected in recent years
  - Both by governments and by private entities
- Data and knowledge extracted by data mining techniques represent a key asset to the society
  - Analyzing trends and patterns
  - Formulating public policies
- Laws and regulations require that some collected data must be made public
  - For example, Census data

## 1.5 Public Data Conundrum

- Health-care datasets
  - Clinical studies, hospital discharge databases
- Genetic datasets
  - \$1000 genome, HapMap, deCode
- Demographic datasets
  - U.S. Census Bureau, sociology studies
- Search logs, recommender systems, social networks, blogs
  - AOL search data, social networks of blogging sites, Netflix movie ratings, Amazon

## 1.6 What About Privacy?

- First thought: anonymize the data
- How?
- Remove “personally identifying information” (PII)
  - Name, Social Security number, phone number, email, address... what else?
  - Anything that identifies the person directly
- Is this enough?

## 1.7 Re-identification by Linking

Microdata					Voter registration data			
ID	QID	SA			Name	Zipcode	Age	Sex
Name	Zipcode	Age	Sex	Disease				
Alice	47677	29	F	Ovarian Cancer	Alice	47677	29	F
Betty	47602	22	F	Ovarian Cancer	Bob	47983	65	M
Charles	47678	27	M	Prostate Cancer	Carol	47677	22	F
David	47905	43	M	Flu	Dan	47532	23	M
Emily	47909	52	F	Heart Disease	Ellen	46789	43	F
Fred	47906	47	M	Heart Disease				

ID is the PII, the rest isn't. When combined they are quasi-identifiers  
Voter registration data is all PII.

## 1.8 Latanya Sweeney's Attack (1997)

### Massachusetts hospital discharge dataset

Medical Data Released as Anonymous

SSN	Name	City	Date Of Birth	Sex	ZIP	Marital Status	Problem
			09/27/64	female	02139	divorced	hypertension
			09/30/64	female	02139	divorced	obesity
	asian		04/15/64	male	02139	married	chest pain
	asian		04/15/64	male	02139	married	obesity
	black		03/13/63	male	02138	married	hypertension
	black		03/18/63	male	02138	married	shortness of breath
	black		09/13/64	female	02141	married	shortness of breath
	black		09/07/64	female	02141	married	obesity
	white		05/14/61	male	02138	single	chest pain
	white		05/08/61	male	02138	single	obesity
	white		09/15/61	female	02142	widow	shortness of breath

Voter List

Name	Address	City	ZIP	DOB	Sex	Party
Sue J. Carlson	1459 Main St.	Cambridge	02142	9/15/61	female	democrat

Figure 1.8- Identifying anonymous data by linking to external data

### Public voter dataset

## 1.9 Quasi-Identifiers

- Key attributes
  - Name, address, phone number - uniquely identifying!
  - Always removed before release
- Quasi-identifiers
  - (5-digit ZIP code, birth date, gender) uniquely identify 87% of the population in the U.S.
  - Can be used for linking anonymized dataset with other datasets

## 1.10 Classification of Attributes

- Sensitive attributes
  - Medical records, salaries, etc.
  - These attributes are what the researchers need, so they are always released directly

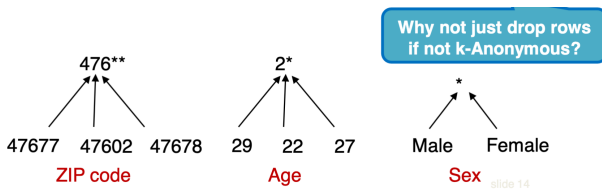
Key Attribute (often PII)	Quasi-identifier			Sensitive attribute
Name	DOB	Gender	Zipcode	Disease
Andre	1/21/76	Male	53715	Heart Disease
Beth	4/13/86	Female	53715	Hepatitis
Carol	2/28/76	Male	53703	Brochitis
Dan	1/21/76	Male	53703	Broken Arm
Ellen	4/13/86	Female	53706	Flu
Eric	2/28/76	Female	53706	Hang Nail

## 1.11 K-Anonymity: Intuition

- The information for each person contained in the released table cannot be distinguished from at least k-1 individuals whose information also appears in the release
  - Example: you try to identify a man in the released table, but the only information you have is his birth date and gender. There are k men in the table with the same birth date and gender.
- Any quasi-identifier present in the released table must appear in at least k records

## 1.12 k-Anonymity via Generalization

- Goal of k-Anonymity
  - Each record is indistinguishable from at least k-1 other records
  - These k records form an equivalence class
- Generalization: replace quasi-identifiers with less specific, but semantically consistent values



## 1.13 Achieving k-Anonymity

- Generalization
  - Replace specific quasi-identifiers with less specific values until get k identical values
  - Partition ordered-value domains into intervals
- Problem: Suppression
  - When generalization causes too much information loss
  - This is common with “outliers”
- Lots of algorithms in the literature
  - Aim to produce “useful” anonymizations
  - ... usually without any clear notion of utility

## 1.14 Example of a k-Anonymous Table

	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

Figure 2 Example of k-anonymity, where k=2 and QI={Race, Birth, Gender, ZIP}

## 1.15 Example of Generalization (1)

Released table

	Race	Birth	Gender	ZIP	Problem
r1	Black	1965	m	0214*	short breath
r2	Black	1965	m	0214*	chest pain
r3	Black	1965	f	0213*	Hypertension
r4	Black	1965	f	0213*	Hypertension
r5	Black	1964	f	0213*	obesity
r6	Black	1964	f	0213*	chest pain
r7	White	1964	m	0213*	chest pain
r8	White	1964	m	0213*	obesity
r9	White	1964	m	0213*	short breath
r10	White	1967	m	0213*	chest pain
r11	White	1967	m	0213*	chest pain

External data Source

Name	Birth	Gender	ZIP	Race
Andre	1964	m	02135	White
Beth	1964	f	55410	Black
Carol	1964	f	90210	White
Dan	1967	m	02174	White
Ellen	1968	f	02237	White

By linking these 2 tables, you still don't learn Andre's problem

## 1.16 Example of Generalization (2)

Microdata

QID			SA
Zipcode	Age	Sex	Disease
47677	29	F	Ovarian Cancer
47602	22	F	Ovarian Cancer
47678	27	M	Prostate Cancer
47905	43	M	Flu
47909	52	F	Heart Disease
47906	47	M	Heart Disease

Generalized table

QID			SA
Zipcode	Age	Sex	Disease
476**	2*	*	Ovarian Cancer
476**	2*	*	Ovarian Cancer
476**	2*	*	Prostate Cancer
4790*	[43,52]	*	Flu
4790*	[43,52]	*	Heart Disease
4790*	[43,52]	*	Heart Disease

Released table is 3-anonymous

Technically yes, but what does k-Anonymity miss?

If the adversary knows Alice's quasi-identifier (47677, 29, F), they still do not know which of the first 3 records corresponds to Alice's record