# 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 ¡AnonID, Query, QueryTime, ItemRank, ClickURL¿
  - 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? No!

## 1.7 Re-identification by Linking

Microdata

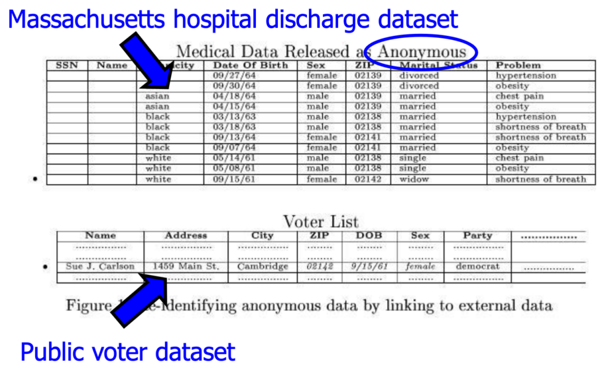
| ID | QID | | | SA |
|---|---|---|---|---|
| Name | Zipcode | Age | Sex | Disease |
| Alice | 47677 | 29 | F | Ovarian Cancer |
| Betty | 47602 | 22 | F | Ovarian Cancer |
| Charles | 47678 | 27 | M | Prostate Cancer |
| David | 47905 | 43 | M | Flu |
| Emily | 47909 | 52 | F | Heart Disease |
| Fred | 47906 | 47 | M | Heart Disease |

Voter registration data

| Name | Zipcode | Age | Sex |
|---|---|---|---|
| Alice | 47677 | 29 | F |
| Bob | 47983 | 65 | M |
| Carol | 47677 | 22 | F |
| Dan | 47532 | 23 | M |
| Ellen | 46789 | 43 | F |

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)




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 | 04/18/64 | 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 …e-identifying anonymous data by linking to external data

## 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

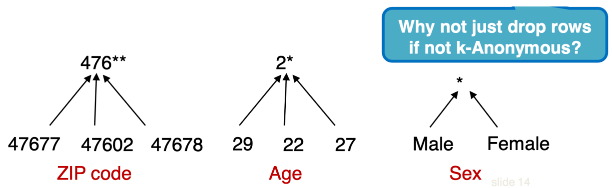
| **Key Attribute (often PII)** | **Quasi-identifier** | | | **Sensitive attribute** |
|---|---|---|---|---|
| Name | DOB | Gender | Zipcode | Disease |
| Andre | 1/21/76 | Male | 53715 | Heart Disease |
| Beth | 4/13/86 | Female | 53715 | Hepatitis |
| Carol | 2/28/76 | Male | 53703 | Brochitis |
| 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
- If you just drop rows if not k-anonymous you cause missingness not at random
- you want to avoid dropping rows to avoid changing decisions, you can just generate attributes instead
- Example: replace 5-digit ZIP code with 3-digit ZIP code




## 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

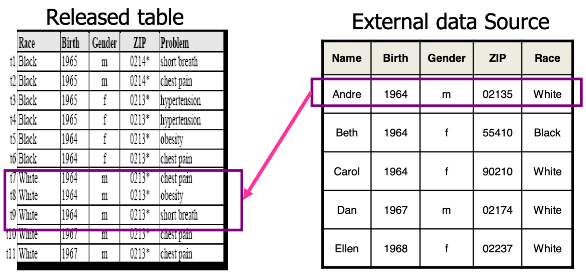
## 1.14 Example of a k-Anonymous Table

| | 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*}

k = 2 and not k = 3 because we take the lowest k value.

## 1.15 Example of Generalization (1)



Released table

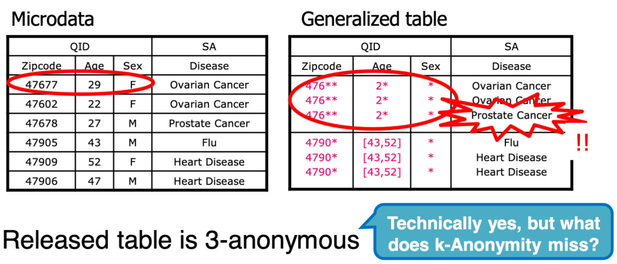
| | Race | Birth | Gender | ZIP | Problem |
|---|---|---|---|---|---|
| t1 | Black | 1965 | m | 0214* | short breath |
| t2 | Black | 1965 | m | 0214* | chest pain |
| t3 | Black | 1965 | f | 0213* | hypertension |
| t4 | Black | 1965 | f | 0213* | hypertension |
| t5 | Black | 1964 | f | 0213* | obesity |
| 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 |

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

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

## 1.17 Curse of Dimensionality

- Generalization fundamentally relies on locality of quasi-identifiers
  - Each record must have k close neighbors
- Real-world datasets are very sparse
  - Many attributes (dimensions)
    - Netflix Prize dataset: 17,000 dimensions
    - Amazon customer records: several million dimensions
  - "Nearest neighbor" is very far
- Projection to low dimensions loses all info
  - k-anonymized datasets are useless

## 1.18 HIPAA Privacy Rule (US)

"Under the safe harbor method, covered entities must remove all of a list of 18 enumerated identifiers and have no actual knowledge that the information remaining could be used, alone or in combination, to identify a subject of the information."

"The identifiers that must be removed include direct identifiers, such as name, street address, social security number, as well as other identifiers, such as birth date, admission and discharge dates, and five-digit zip code. The safe harbor requires removal of geographic subdivisions smaller than a State, except for the initial three digits of a zip code if the geographic unit formed by combining all zip codes with the same initial three digits contains more than 20,000 people. In addition, age, if less than 90, gender, ethnicity, and other demographic information not listed may remain in the information. The safe harbor is intended to provide covered entities with a simple, definitive method that does not require much judgment by the covered entity to determine if the information is adequately de-identified."

## 1.19 Two (and a Half) Interpretations

1. Membership disclosure: Attacker cannot tell that a given person in the dataset
2. Sensitive attribute disclosure: Attacker cannot tell that a given person has a certain sensitive attribute
3. Identity disclosure: Attacker cannot tell which record corresponds to a given person

This (3) interpretation is correct, assuming the attacker does not know anything other than quasi-identifiers But this does not imply any privacy! Example: k clinical records, all HIV+

## 1.20 Attacks on k-Anonymity

- k-Anonymity does not provide privacy if
  - Sensitive values in an equivalence class lack diversity
  - The attacker has background knowledge

Homogeneity attack

| Bob | |
|---|---|
| ***Zipcode*** | ***Age*** |
| 47678 | 27 |

Background knowledge attack

| Carl | |
|---|---|
| ***Zipcode*** | ***Age*** |
| 47673 | 36 |

A 3-anonymous patient table

| Zipcode | Age | Disease |
|---|---|---|
| 476** | 2* | Heart Disease |
| 476** | 2* | Heart Disease |
| 476** | 2* | Heart Disease |
| 4790* | ≥40 | Flu |
| 4790* | ≥40 | Heart Disease |
| 4790* | ≥40 | Cancer |
| 476** | 3* | Heart Disease |
| 476** | 3* | Cancer |
| 476** | 3* | Cancer |

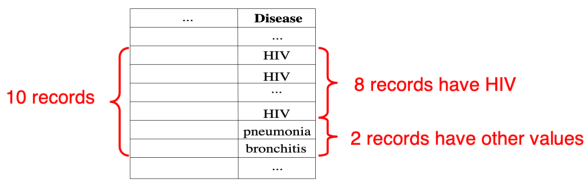
## 1.21 l-Diversity

| | | |
|---|---|---|
| Caucas | 787XX | Flu |
| Caucas | 787XX | Shingles |
| Caucas | 787XX | Acne |
| Caucas | 787XX | Flu |
| Caucas | 787XX | Acne |
| Caucas | 787XX | Flu |
| Asian/AfrAm | 78XXX | Flu |
| Asian/AfrAm | 78XXX | Flu |
| Asian/AfrAm | 78XXX | Acne |
| Asian/AfrAm | 78XXX | Shingles |
| Asian/AfrAm | 78XXX | Acne |
| Asian/AfrAm | 78XXX | Flu |

Sensitive attributes must be "diverse" within each quasi-identifier equivalence class

## 1.22 Distinct l-Diversity

- Each equivalence class of quasi-identifiers has at least l well-represented sensitive values
- Doesn't prevent probabilistic inference attacks




I = 10% here.

## 1.23 Other Versions of l-Diversity

- Probabilistic l-diversity
  - The frequency of the most frequent value in an equivalence class is bounded by 1/l
- Entropy l-diversity
  - The entropy of the distribution of sensitive values in each equivalence class is at least log(l)
- Recursive (c,l)-diversity
  - $r_1 < c(r_l + r_{l+1} + \cdots + r_m)$, where $r_i$ is the frequency of the $i$-th most frequent value
  - Intuition: the most frequent value does not appear too frequently

## 1.24 t-Closeness

(not tested) Distribution of sensitive attributes within each quasi-identifier group should be "close" to their distribution in the entire original database

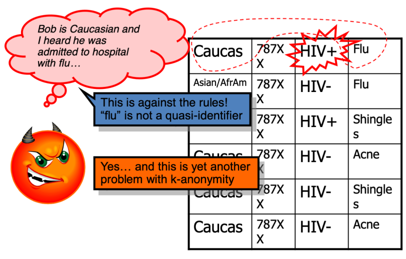
| | | |
|---|---|---|
| Caucas | 787XX | Flu |
| Caucas | 787XX | Shingles |
| Caucas | 787XX | Acne |
| Caucas | 787XX | Flu |
| Caucas | 787XX | Acne |
| Caucas | 787XX | Flu |
| Asian/AfrAm | 78XXX | Flu |
| Asian/AfrAm | 78XXX | Flu |
| Asian/AfrAm | 78XXX | Acne |
| Asian/AfrAm | 78XXX | Shingles |
| Asian/AfrAm | 78XXX | Acne |
| Asian/AfrAm | 78XXX | Flu |

## 1.25 Anonymous, "t-Close" Dataset

This is k-anonymous, l-diverse and t-close... ...so secure, right?

| | | | |
|---|---|---|---|
| Caucas | 787X X | HIV+ | Flu |
| Asian/AfrAm | 787X X | HIV- | Flu |
| Asian/AfrAm | 787X X | HIV+ | Shingle s |
| Caucas | 787X X | HIV- | Acne |
| Caucas | 787X X | HIV- | Shingle s |
| Caucas | 787X X | HIV- | Acne |

## 1.26 What Does Attacker Know?




## 1.27 k-Anonymity is Not Enough!

- Syntactic
  - Focuses on data transformation, not on what can be learned from the anonymized dataset
  - “k-anonymous” dataset can leak sensitive information
- “Quasi-identifier” fallacy
  - Assumes a priori that attacker will not know certain information about their target
- Can increase levels of anonymity, but ...
  - Destroys utility of many real-world datasets

## 1.28 Exam question

How can you attack a k-anonymous dataset?