# Database Privacy: Principles and Algorithms (Part I)

Zhenjie Zhang

Advanced Digital Sciences Center, Singapore

（Thanks to Marianne Winslett, Xiaokui Xiao, Gerome Miklau, Yin Yang and others for contributing slides)

# Outline

- Why privacy?
- Privacy Attacking Examples
- Conventional principles and limitations
  - K-anonymity
  - L-diversity
  - T-closeness

# On the Internet, nobody knows you're a dog?


*The New Yorker*, July 5, 1993

Your personal information is kept by

- Government agencies
- Banks/Financial business
- Online shopping web sites
- Advertising companies

# Publishing sensitive data about individuals.

- Medical research
  - What treatments have the best outcomes?
  - How can we recognize the onset of disease earlier?
  - Are certain drugs better for certain phenotypes?
- Web search
  - What are people really looking for when they search?
  - How can we give them the most authoritative answers?
- Public health
  - Where are our outbreaks of unpleasant diseases?
  - What behavior patterns or patient characteristics are correlated with these diseases?

# Publishing sensitive data about individuals.

- Social and computer networking
  - What is the pattern of phone/data/multimedia network usage? How can we better use existing (or plan new) infrastructure to handle this traffic?
  - How do people relate to one another, e.g., as mediated by Facebook?
  - How is society evolving (Census data)?
- Industrial data (individual = company; need SMC if no TTP)
  - What were the total sales, over all companies, in a sector last year/quarter/month?
  - What were the characteristics of those sales: who were the buyers, how large were the purchases, etc.?

# Today, access to these data sets is usually strictly controlled.

Only available:

- Inside the company/agency that collected the data
- Or after signing a legal contract
  - Click streams, taxi data
- Or in very coarse-grained summaries
  - Public health
- Or after a very long wait
  - US Census data details
- Or with definite privacy issues
  - US Census reports, the AOL click stream, dbGaP summary tables, Enron email
- Or with IRB (Institutional Review Board) approval
  - dbGaP summary tables

Society would benefit if we could publish some useful form of the data, without having to worry about privacy.

# Why is access so strictly controlled?

No one should learn who had which disease.

| Name | Age | Sex | Zipcode | Disease |
|---|---|---|---|---|
| Andy | 5 | M | 12000 | gastric ulcer |
| Bill | 9 | M | 14000 | dyspepsia |
| Ken | 6 | M | 18000 | pneumonia |
| Nash | 8 | M | 19000 | bronchitis |
| Joe | 12 | M | 22000 | pneumonia |
| Sam | 19 | M | 24000 | pneumonia |
| Linda | 21 | F | 58000 | flu |
| Jane | 26 | F | 36000 | gastritis |
| Sarah | 28 | F | 37000 | pneumonia |
| Mary | 56 | F | 33000 | flu |

“**Microdata**”

# What if we "de-identify" the records by removing names?

| Name | Age | Sex | Zipcode | Disease |
|---|---|---|---|---|
| Andy | 5 | M | 12000 | gastric ulcer |
| Bill | 9 | M | 14000 | dyspepsia |
| Ken | 6 | M | 18000 | pneumonia |
| Nash | 8 | M | 19000 | bronchitis |
| Joe | 12 | M | 22000 | pneumonia |
| Sam | 19 | M | 24000 | pneumonia |
| Linda | 21 | F | 58000 | flu |
| Jane | 26 | F | 36000 | gastritis |
| Sarah | 28 | F | 37000 | pneumonia |
| Mary | 56 | F | 33000 | flu |

publish

| Age | Sex | Zipcode | Disease |
|---|---|---|---|
| 5 | M | 12000 | gastric ulcer |
| 9 | M | 14000 | dyspepsia |
| 6 | M | 18000 | pneumonia |
| 8 | M | 19000 | bronchitis |
| 12 | M | 22000 | pneumonia |
| 19 | M | 24000 | pneumonia |
| 21 | F | 58000 | flu |
| 26 | F | 36000 | gastritis |
| 28 | F | 37000 | pneumonia |
| 56 | F | 33000 | flu |

# We can re-identify people, absolutely or probabilistically

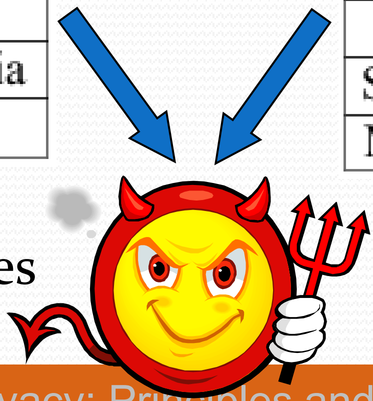
The published table

| Age | Sex | Zipcode | Disease |
|---|---|---|---|
| 5 | M | 12000 | gastric ulcer |
| 9 | M | 14000 | dyspepsia |
| 6 | M | 18000 | pneumonia |
| 8 | M | 19000 | bronchitis |
| 12 | M | 22000 | pneumonia |
| 19 | M | 24000 | pneumonia |
| 21 | F | 58000 | flu |
| 26 | F | 36000 | gastritis |
| 28 | F | 37000 | pneumonia |
| 56 | F | 33000 | flu |

Quasi-identifier (QI) attributes

A voter registration list

| Name | Age | Sex | Zipcode |
|---|---|---|---|
| Andy | 5 | M | 12000 |
| Bill | 9 | M | 14000 |
| Ken | 6 | M | 18000 |
| Nash | 8 | M | 19000 |
| *Mike* | *7* | *M* | *17000* |
| Joe | 12 | M | 22000 |
| Sam | 19 | M | 24000 |
| Linda | 21 | F | 58000 |
| Jane | 26 | F | 36000 |
| Sarah | 28 | F | 37000 |
| Mary | 56 | F | 33000 |



**"Background knowledge"**

# 87% of Americans can be uniquely identified by {zip code, gender, date of birth}.

actually 63%
[Golle 06]

Latanya Sweeney [*International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 2002] used this approach to re-identify the medical record of an ex-governor of Massachusetts.

# Outline

- Why privacy?
- Privacy Attacking Examples
- Conventional Principles and limitations
  - K-anonymity
  - L-diversity
  - T-closeness

# Real query logs can be very useful to CS researchers.  But click history can uniquely identify a person.

<*AnonID, Query, QueryTime, ItemRank, domain name clicked*>

What the New York Times did:

- Find all log entries for AOL user 4417749
- Multiple queries for businesses and services in Lilburn, GA (population 11K)
- Several queries for Jarrett Arnold
  - Lilburn has 14 people with the last name Arnold
- NYT contacts them, finds out AOL User 4417749 is Thelma Arnold

# Just because data looks hard to re-identify, doesn't mean it *is*.

[Narayanan and Shmatikov, Oakland 08]

In 2009, the Netflix movie rental service offered a $1,000,000 prize for improving their movie recommendation service.

| | High School Musical 1 | High School Musical 2 | High School Musical 3 | Twilight |
|---|---|---|---|---|
| Customer #1 | 4 | 5 | 5 | ? |

Training data: ~100M ratings of 18K movies from ~500K randomly selected customers, plus dates

Only 10% of their data; slightly perturbed

# We can re-identify a Netflix rater if we know just a little bit about her (from life, IMDB ratings, blogs, …).

- 8 movie ratings (≤ 2 wrong, dates ±2 weeks) ➔ re-identify 99% of raters
- 2 ratings, ±3 days ➔ re-identify 68% of raters
  - Relatively few candidates for the other 32% (especially with movies outside the top 100)
- Even a handful of IMDB comments allows Netflix re-identification, in many cases
  - 50 IMDB users → re-identify 2 with very high probability, one from ratings, one from dates

# Why should we care about this innocuous data set?

- *All* movie ratings → political and religious opinions, sexual orientation, …
- *Everything* bought in a store → private life details
- *Every* doctor visit → private life details

"One customer … sued Netflix, saying she thought her rental history could reveal that she was a lesbian before she was ready to tell everyone."

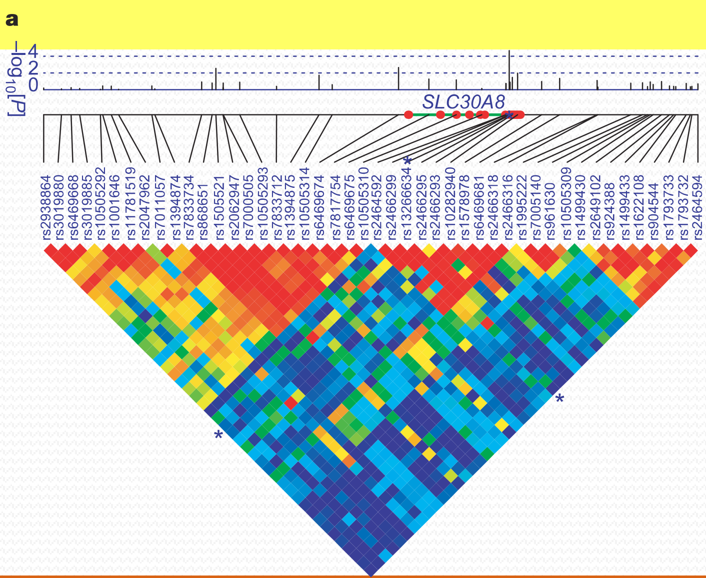
# It is becoming routine for medical studies to include a genetic component.

Genome-wide association studies (GWAS) aim to identify the correlation between diseases, e.g., diabetes, and the patient's DNA, by comparing people with and without the disease.

GWAS papers usually include detailed correlation statistics.

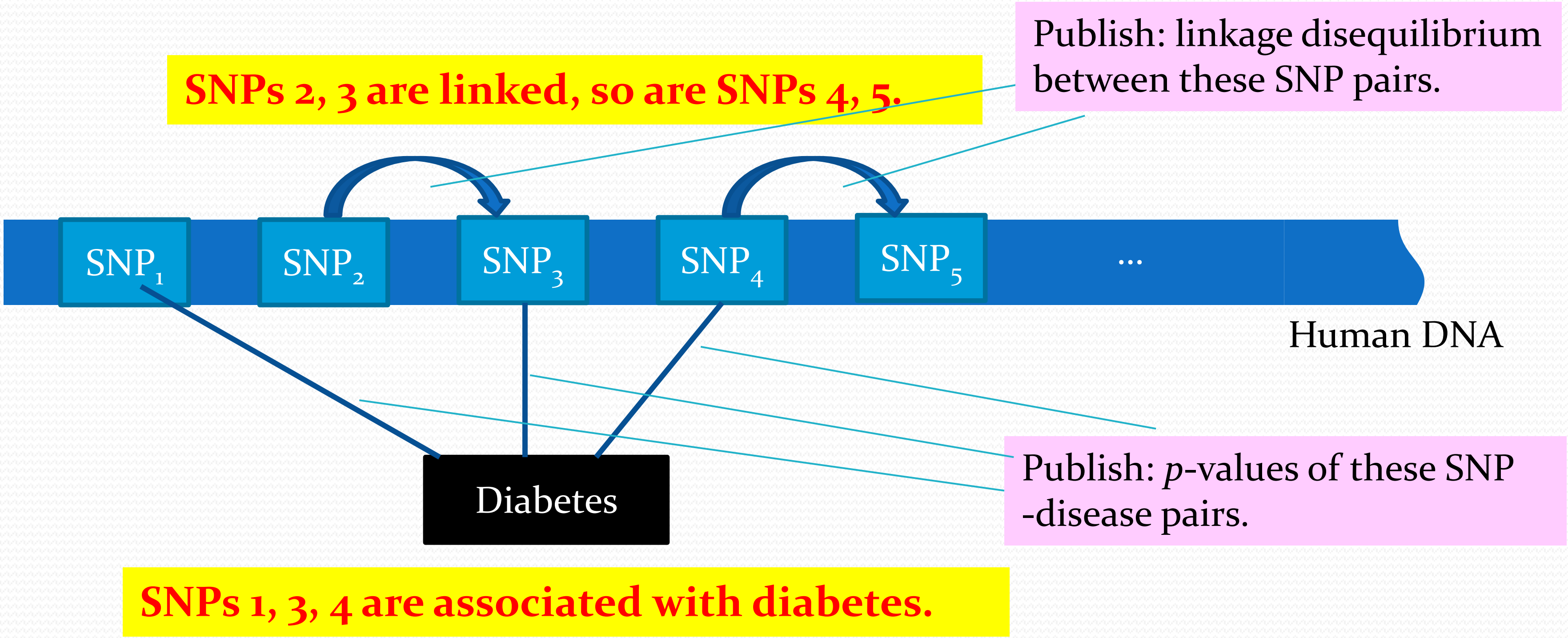
**Our attack**: uncover the identities of the patients in a GWAS

- For studies of up to moderate size, a significant fraction of people, determine whether a specific person has participated in a particular study within 10 seconds, with high confidence!




**A genome-wide association study identifies novel risk loci for type 2 diabetes, Nature 445, 881-885 (22 February 2007)**

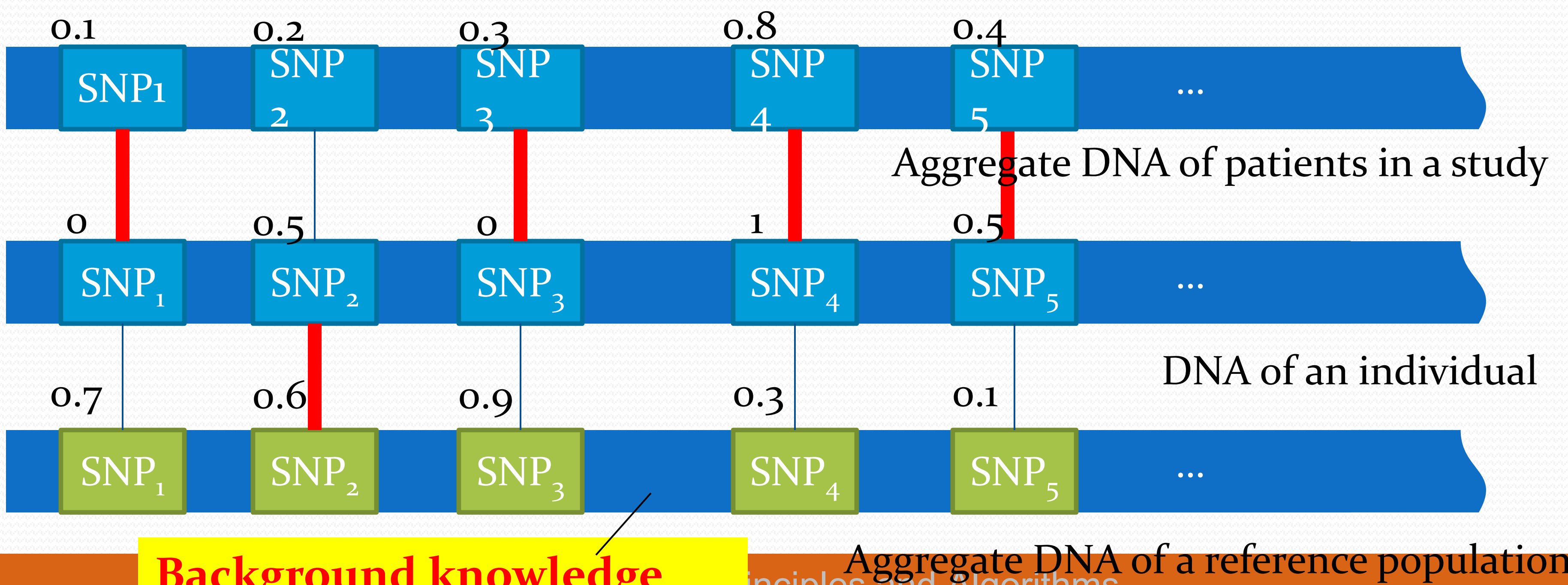
# GWAS papers usually include detailed correlation statistics.



# Privacy attacks can use SNP-disease association.

Idea [Homer et al. *PloS Genet.*'08, Jacobs et al. *Nature*'09]:

- Obtain aggregate SNP info from the published *p*-values (1)
- Obtain a sample DNA of the target individual (2)
- Obtain the aggregate SNP info of a ref. population (3)
- Compare (1), (2), (3)



# Outline

- Why privacy?
- Privacy Attacking Examples
- Conventional Principles and limitations
  - K-anonymity
  - L-diversity
  - T-closeness

# Issues

- Privacy principle
  - What is adequate privacy protection?
- Distortion approach
  - How can we achieve the privacy principle,
  - while maximizing the utility of the data?

# Different applications may have different privacy protection needs.

**Membership disclosure**: Attacker cannot tell that a given person is/was in the data set (e.g., a set of AIDS patient records or the summary data from a data set like dbGaP).

- δ-presence [Nergiz et al., 2007].
- Differential privacy [Dwork, 2007].

**Sensitive attribute disclosure**: Attacker cannot tell that a given person has a certain sensitive attribute.

- *l*-diversity [Machanavajjhala et al., 2006].
- *t*-closeness [Li et al., 2007].

**Identity disclosure**: Attacker cannot tell which record corresponds to a given person.

- *k*-anonymity [Sweeney, 2002].

# Privacy principle 1: *k*-anonymity

your quasi-identifiers are indistinguishable from ≥ *k* other people's.

[Sweeney, *Int'l J. on Uncertainty, Fuzziness and Knowledge-based Systems*, 2002]

## 2-anonymous generalization:

A voter registration list

| Name | Age | Sex | Zipcode |
|---|---|---|---|
| Andy | 5 | M | 12000 |
| Bill | 9 | M | 14000 |
| Ken | 6 | M | 18000 |
| Nash | 8 | M | 19000 |
| *Mike* | *7* | *M* | *17000* |
| Joe | 12 | M | 22000 |
| Sam | 19 | M | 24000 |
| Linda | 21 | F | 58000 |
| Jane | 26 | F | 36000 |
| Sarah | 28 | F | 37000 |
| Mary | 56 | F | 33000 |

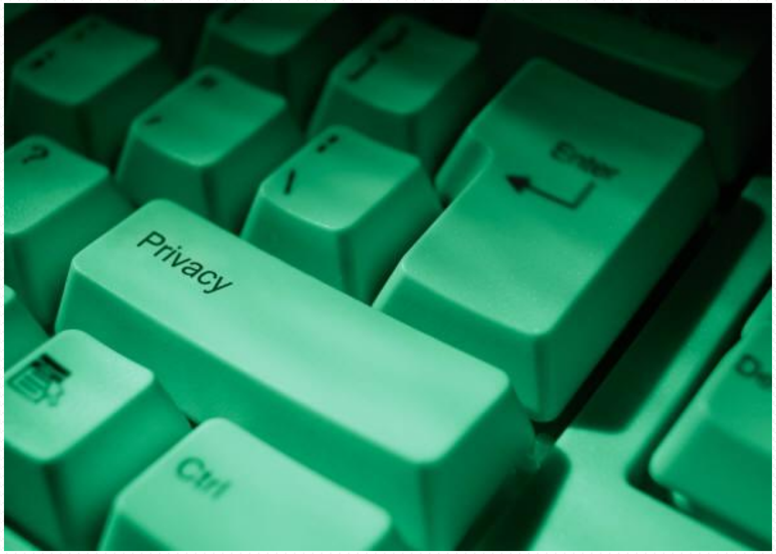
Sensitive attribute

QI attributes

4 QI groups

| Age | Sex | Zipcode | Disease |
|---|---|---|---|
| [1, 10] | M | [10001, 15000] | gastric ulcer |
| [1, 10] | M | [10001, 15000] | dyspepsia |
| [1, 10] | M | [15001, 20000] | pneumonia |
| [1, 10] | M | [15001, 20000] | bronchitis |
| [11, 20] | M | [20001, 25000] | pneumonia |
| [11, 20] | M | [20001, 25000] | pneumonia |
| [21, 60] | F | [30000, 60000] | flu |
| [21, 60] | F | [30000, 60000] | gastritis |
| [21, 60] | F | [30000, 60000] | pneumonia |
| [21, 60] | F | [30000, 60000] | flu |

# The biggest advantage of k-anonymity is that people can understand it.



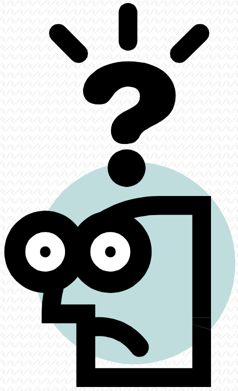

And often it can be computed fast.

But in general, it is easy to attack.

# *k*-anonymity . . . or how <u>not</u> to define privacy.

[Shmatikov]

- Does not say anything about the computations to be done on the data (utility).
- Assumes that attacker will be able to join <u>only</u> on quasi-identifiers.



Intuitive reasoning:

- *k*-anonymity prevents attacker from telling which record corresponds to which person.
- Therefore, attacker cannot tell that a certain person has a particular value of a sensitive attribute.

**This reasoning is fallacious!**

# *k*-anonymity does not provide privacy if the sensitive values in an equivalence class lack diversity, or the attacker has certain background knowledge.

From a voter registration list

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 |

# Updates can also destroy k-anonymity.

What is Joe's disease? Wait for his birthday.

A voter registration list
plus dates of birth (not shown)

| Name | Age | Sex | Zipcode |
|---|---|---|---|
| Andy | 5 | M | 12000 |
| Bill | 9 | M | 14000 |
| Ken | 6 | M | 18000 |
| Nash | 8 | M | 19000 |
| *Mike* | *7* | *M* | *17000* |
| Joe | 10 | M | 17000 |
| Sam | 19 | M | 24000 |
| Linda | 21 | F | 58000 |
| Jane | 26 | F | 36000 |
| Sarah | 28 | F | 37000 |
| Mary | 56 | F | 33000 |

No "diversity" in this QI group.

| Age | Sex | Zipcode | Disease |
|---|---|---|---|
| [1, 10] | M | [10001, 15000] | gastric ulcer |
| [1, 10] | M | [10001, 15000] | dyspepsia |
| [1, 10] | M | [15001, 20000] | pneumonia |
| [1, 10] | M | [15001, 20000] | bronchitis |
| [11, 20] | M | [20001, 25000] | pneumonia |
| [11, 20] | M | [20001, 25000] | pneumonia |
| [21, 60] | F | [30000, 60000] | flu |
| [21, 60] | F | [30000, 60000] | gastritis |
| [21, 60] | F | [30000, 60000] | pneumonia |
| [21, 60] | F | [30000, 60000] | flu |

# Principle 2: *l*-diversity

[Machanavajjhala et al., *ICDE*, 2006]

Each QI group should have at least *l* "well-represented" sensitive values.

# Maybe each QI-group must have *l different* sensitive values?

A 2-diverse table

| Age | Sex | Zipcode | Disease |
|---|---|---|---|
| [1, 5] | M | [10001, 15000] | gastric ulcer |
| [1, 5] | M | [10001, 15000] | dyspepsia |
| [6, 10] | M | [15001, 20000] | pneumonia |
| [6, 10] | M | [15001, 20000] | bronchitis |
| [11, 20] | F | [20001, 25000] | flu |
| [11, 20] | F | [20001, 25000] | pneumonia |
| [21, 60] | F | [30001, 60000] | gastritis |
| [21, 60] | F | [30001, 60000] | gastritis |
| [21, 60] | F | [30001, 60000] | flu |
| [21, 60] | F | [30001, 60000] | flu |

# We can attack this probabilistically.

If we know Joe's QI group, what is the probability he has HIV?

| … | **Disease** |
|---|---|
| | … |
| | HIV |
| | HIV |
| | … |
| | HIV |
| | pneumonia |
| | bronchitis |
| | … |

A QI group with 100 tuples

98 tuples

The conclusion researchers drew: The most frequent sensitive value in a QI group cannot be too frequent.

# Even then, we can still attack using background knowledge.

Joe has HIV.

Sally knows Joe does not have pneumonia.

Sally can guess that Joe has HIV.

| ... | **Disease** |
|---|---|
| | ... |
| | HIV |
| | ... |
| | HIV |
| | pneumonia |
| | ... |
| | pneumonia |
| | bronchitis |
| | ... |

A QI group with 100 tuples

50 tuples

49 tuples

# *l*-diversity variants have been proposed to address these weaknesses.

- 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
  - The most frequent value does not appear too frequently
  - $r_1<c(r_l+r_{l+1}+\ldots+r_m)$, where $r_i$ is the frequency of the *i*-th most frequent value.

# l-diversity can be overkill or underkill.

Original data

| | |
|---|---|
| … | Cancer |
| … | Cancer |
| … | Cancer |
| … | Flu |
| … | Cancer |
| … | Cancer |
| … | Cancer |
| … | Cancer |
| … | Cancer |
| … | Cancer |
| … | Flu |
| … | Flu |

99% have cancer

Anonymization A

| | |
|---|---|
| Q1 | Flu |
| Q1 | Flu |
| Q1 | Cancer |
| Q1 | Flu |
| Q1 | Cancer |
| Q1 | Cancer |
| Q2 | Cancer |
| Q | |

Anonymization B

| | |
|---|---|
| Q1 | Flu |
| Q1 | Cancer |
| Q1 | Cancer |
| Q1 | Cancer |
| Q1 | Cancer |
| Q1 | Cancer |
| Q2 | Cancer |
| Q2 | Flu |

99% cancer ⇒ quasi-identifier group is <u>not</u> “diverse”, yet anonymized database does not leak much new info.

50% cancer ⇒ quasi-identifier group is “diverse”
**This leaks a ton of new information**

**Diversity does not *inherently* benefit privacy.**

# Principle 3: t-Closeness

[Li et al. ICDE '07]

| | | |
|---|---|---|
| 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 |

Distribution of sensitive attributes within each quasi-identifier group should be "close" to their distribution in the entire original DB

# Then we can bound the knowledge that the attacker gains by seeing a particular anonymization.

## Adversarial belief

| Belief | Knowledge |
|---|---|
| $B_0$ | External Knowledge |
| $B_1$ | Overall distribution of sensitive values |
| $B_2$ | Distribution of sensitive values in a particular group |

Released table

| Age | Zip code | ...... | Gender | Disease |
|---|---|---|---|---|
| 2* | 479** | ...... | Male | Flu |
| 2* | 479** | ...... | Male | Heart Disease |
| 2* | 479** | ...... | Male | Cancer |
| . | . | ...... | . | . |
| . | . | ...... | . | . |
| . | . | ...... | . | . |
| ≥50 | 4766* | ...... | * | Gastritis |

**Only applicable when we can define the distance between values, e.g., using a hierarchy of diagnoses.**

# How anonymous is this 4-anonymous, 3-diverse, and perfectly-*t*-close data?

| | | | |
|---|---|---|---|
| Asian/AfrAm | 787XX | HIV- | Acne |
| Asian/AfrAm | 787XX | HIV- | Acne |
| Asian/AfrAm | 787XX | HIV- | Flu |
| Asian/AfrAm | 787XX | HIV+ | Shingles |
| Caucasian | 787XX | HIV+ | Flu |
| Caucasian | 787XX | HIV- | Acne |
| Caucasian | 787XX | HIV- | Shingles |
| Caucasian | 787XX | HIV- | Acne |

# That depends on the attacker's background knowledge.

*My coworker Bob's shingles got so bad that he is in the hospital. He looks Asian to me…*

This is against the rules, because flu is not a quasi-identifier.

| | | | |
|---|---|---|---|
| Asian/AfrAm | 787XX | HIV- | Acne |
| Asian/AfrAm | 787XX | HIV- | Acne |
| Asian/AfrAm | 787XX | HIV- | Flu |
| Asian/AfrAm | 787XX | HIV+ | Shingles |
| Caucasian | 787XX | HIV+ | Flu |
| Caucasian | 787XX | HIV- | Acne |
| Caucasian | 787XX | HIV- | Shingles |
| Caucasian | 787XX | HIV- | Acne |

In the real world, almost *anything* could be personally identifying (as we saw with Netflix).

# There are probably 100 other related proposed privacy principles…

- $k$-gather, $(a, k)$-anonymity, personalized anonymity, positive disclosure-recursive $(c, l)$-diversity, non-positive-disclosure $(c_1, c_2, l)$-diversity, $m$-invariance, $(c, t)$-isolation, …

And for other data models, e.g., graphs:

- $k$-degree anonymity, $k$-neighborhood anonymity, $k$-sized grouping, $(k, l)$ grouping, …

# … and they suffer from related problems. [Shmatikov]

~~Trying to achieve "privacy" by syntactic transformation of the data~~

- ~~Scrubbing of PII, k-anonymity, l-diversity…~~

**Fatally flawed!**

- Insecure against attackers with arbitrary background info
- Do not compose (anonymize twice $\Rightarrow$ reveal data)
- No meaningful notion of privacy
- No meaningful notion of utility

*Does he go too far?*

# And there is an impossibility result that applies to all of them.

[Dwork, Naor 2006]

For any reasonable definition of "privacy breach" and "sanitization", with high probability some adversary can breach some sanitized DB.

Example:

- Private fact: my exact height
- Background knowledge: I'm 5 inches taller than the average American woman
- San(DB) allows computing average height of US women
- This breaks my privacy … even if my record is <u>not</u> in the database!