CS489/698 Privacy, Cryptography, Network and Data Security

Syntactic Notions of Privacy

Spring 2024, Monday/Wednesday 11:30am-12:50pm

A Recap on Linking Attacks

- As the name suggests, linking attacks find connections between two different sources of leakage that, alone, seem harmless.
- Famous example, from [1]:

The Group Insurance Comission (GIC) in Massachusetts, sold data from 135,000 state employees to industry and researchers. They believed it was anonymous, so it was fine.

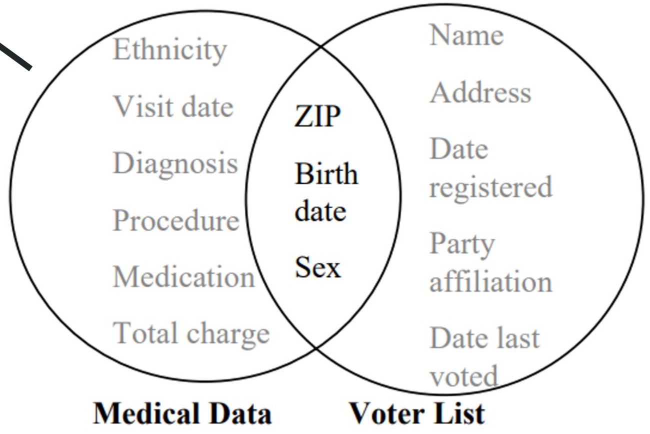


Figure 1 Linking to re-identify data

For $20, you can purchase the voter registration list for Cambridge, Massachusetts

Fun fact: 87% (216 million of 248 million) of the population in the United States had reported characteristics that likely made them **unique** based only on {5-digit ZIP, gender, date of birth}

[1] Sweeney, Latanya. "k-anonymity: A model for protecting privacy." *International journal of uncertainty, fuzziness and knowledge-based systems* 10.05 (2002): 557-570.

Inference across multiple sources

- The inference problem is more severe when the adversary has access to multiple data sources as long as they can link and aggregate the information from different sources

- **Q:** Where do you get these external data sources?

Inference across multiple sources

- The inference problem is more severe when the adversary has access to multiple data sources as long as they can link and aggregate the information from different sources

- **Q:** Where do you get these external data sources?
    - Use publicly available data, e.g. census data, regional records.
    - Purchase data records from a data broker.
    - Governments might also share their dossiers with each other.
    - Large companies may collect information about their customers.

Inference across multiple sources

- Now, what can we learn from combining these datasets that we didn't learn before?
- If these datasets include identifiers that are verinyms, or persistent pseudonyms, one can link data records across these datasets to learn more information about an individual or an entity.

Inference across multiple sources

- Now, what can we learn from combining these datasets that we didn't learn before?
- If these datasets include identifiers that are verinyms, or persistent pseudonyms, one can link data records across these datasets to learn more information about an individual or an entity.
- **Q:** I erased all the identification information before I publicly release the data, would that break the link?

Inference across multiple sources

- Now, what can we learn from combining these datasets that we didn't learn before?
- If these datasets include identifiers that are verinyms, or persistent pseudonyms, one can link data records across these datasets to learn more information about an individual or an entity.
- **Q:** I erased all the identification information before I publicly release the data, would that break the link?
  - We will see a series of inference attacks on public data releases that are supposed to protect the privacy of the data suppliers but failed.

Anonymity failure: AOL Search Data Set

- August 6, 2006: AOL released 20 million search queries from 658,000 users over a 3-month period in 2006.
- AOL assigned a random number to each user:
    - 4417749 “numb fingers”
    - 4417749 “60 single men”
    - 4417749 “landscapers in Lilburn, GA”
    - 4417749 “dog that urinates on everything”
    - 711391 “life in Alaska”
- August 9: New York Times article re-identified user 4417749

Anonymity failure: AOL Search Data Set

- August 6, 2006: AOL released 20 million search queries from 658,000 users over a 3-month period in 2006.
- AOL assigned a random number to each user:
  - 4417749 “numb fingers”
  - 4417749 “60 single men”
  - 4417749 “landscapers in Lilburn, GA”
  - 4417749 “dog that urinates on everything”
  - 711391 “life in Alaska”
- August 9: New York Times article re-identified user 4417749
  - Thelma Arnold, 62-year old widow from Lilburn, GA

Takeaway: simply attaching a random number to each users’ record is insufficient to get a high degree of anonymity.

Anonymity failure: NYC Taxi dataset release

- NYC Taxi Commission released 173 million “anonymized” NYC Taxi trip logs due to a FOIA request
- Each trip log includes information about the trip as well as persistent pseudonyms for each taxi itself”
  - pick-up location (latitude, longitude) and time
  - drop-off location (latitude, longitude) and time
  - MD5 hash of the taxi medallion number
  - MD5 hash of the driver license number
- Parameters collected to learn about taxi usage and traffic patterns.

Anonymity failure: NYC Taxi dataset release

- **Anonymity problem 1** with this data release: Pick-up / drop-off times and locations can be correlated with celebrities' travels (background knowledge from other news sources)
- Example:

  - You know that a celebrity was spotted leaving the JFK airport at 6pm.

    ⇒ You look for pick-up records near JFK at 6pm and see where they drop-off.

    ⇒ After filter out infeasible locations, you might be able to identify the taxi that they took and deduce where they lived or visited.

Anonymity failure: NYC Taxi dataset release

- **Anonymity problem 1** with this data release: Pick-up / drop-off times and locations can be correlated with celebrities' travels (background knowledge from other news sources)
- Example:
  - You know that a celebrity was spotted leaving the JFK airport at 6pm.

    ⇒ You look for pick-up records near JFK at 6pm and see where they drop-off.

    ⇒ After filter out infeasible locations, you

that they took and deduce where they lived or v

Takeaway: Perhaps these drop-offs/pick-ups could be published at a lower granularity, at the cost of lower utility for statistical analysis of traffic, etc.

Anonymity failure: NYC Taxi dataset release

- **Anonymity problem 2** with this data release: Does hashing help with hiding identities of the drivers and taxicabs?
- **Background info:** These two identifiers have the following structure:
  - License numbers are 6 or 7-digit numbers
  - Medallion numbers are either:
    - [0-9][A-Z][0-9][0-9]
    - [A-Z][A-Z][0-9][0-9][0-9]
    - [A-Z][A-Z][A-Z][0-9][0-9][0-9]

Anonymity failure: NYC Taxi dataset release

- **Anonymity problem 2** with this data release: Does hashing help with hiding identities of the drivers and taxicabs?
- **Background info:** These two identifiers have the following structure:
  - License numbers are 6 or 7-digit numbers
  - Medallion numbers are either:
    - [0-9][A-Z][0-9][0-9]
    - [A-Z][A-Z][0-9][0-9][0-9]
    - [A-Z][A-Z][A-Z][0-9][0-9][0-9]

Q: How would you uncover their identities?

Anonymity failure: NYC Taxi dataset release

- **Anonymity problem 2** with this data release: Does hashing help with hiding identities of the drivers and taxicabs?
- **Background info:** These two identifiers have the following structure:
  - License numbers are 6 or 7-digit numbers
  - Medallion numbers are either:
    - [0-9][A-Z][0-9][0-9]
    - [A-Z][A-Z][0-9][0-9][0-9]
    - [A-Z][A-Z][A-Z][0-9][0-9][0-9]

Q: How would you uncover their identities?

A: Brute-force! There are only 1 million license numbers at most, and 17 million medallion numbers

Anonymity failure: NYC Taxi dataset release

- **Anonymity problem 2** with this data release: Does hashing help with hiding identities of the drivers and taxicabs?
- **Background info:** These two identifiers have the following structure:
  - License numbers are 6 or 7-digit numbers
  - Medallion numbers are either:
    - [0-9][A-Z][0-9][0-9]
    - [A-Z][A-Z][0-9][0-9][0-9]
    - [A-Z][A-Z][A-Z][0-9][0-9][0-9]

Q: How would you uncover their identities?

A: Brute-force! There are only 1 million license numbers at most, and 17 million medallion numbers

Takeaway: Hashing identifiers does not provide anonymity. Dictionary attacks are efficient for small input spaces

Anonymity failure: Massachusetts Insurance Health Records

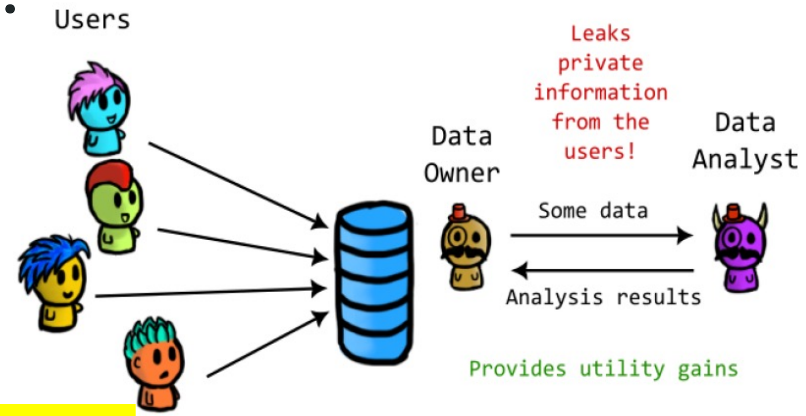
- Massachusetts released “anonymized” health records:
  - ZIP code
  - Gender
  - Date of birth
  - Health information
- Massachusetts’ voter registration list:
  - ZIP code
  - Gender
  - Date of birth
  - Name

Lessons Learned

- Datasets included data that was useful for research (primary data), as well as some identifiers (“quasi-identifiers”).
- “**Quasi-identifiers**” can be used to link data across multiple records in the same dataset (NYC Taxi dataset or AOL search data) or across different datasets (Massachusetts case).
- **Background knowledge** relating to the primary data, can be used to further de-anonymize records.

Moving towards Defences

- We saw many attacks.
- Now, we're going to see some defences.
- How do we measure privacy?
  - **Empirically**:
    - by measuring the performance of an attack
  - **Theoretically**:
    - **Syntactic** notions: measuring a property on the released data / leakage.
    - **Semantic** notions: ensuring the data release mechanism itself has a property (independent of its inputs/outputs)



Syntactic Privacy in relational databases

- Syntactic notions of privacy define a property that the released data must satisfy.
- The notions we will see refer to tabular data (relational databases).
- When talking about a table, the columns are the attributes, and the rows are the data entries or samples.

Syntactic Privacy in relational databases

- The attributes of a table can be classified into:
  - Identifiers: uniquely identify a participant
  - **Quasi-identifiers**: in combination with external information, can identify a participant (ZIP, DOB, Gender, etc.)
  - **Confidential attributes**: contain privacy-sensitive information
  - Non-confidential attributes: are not considered sensitive

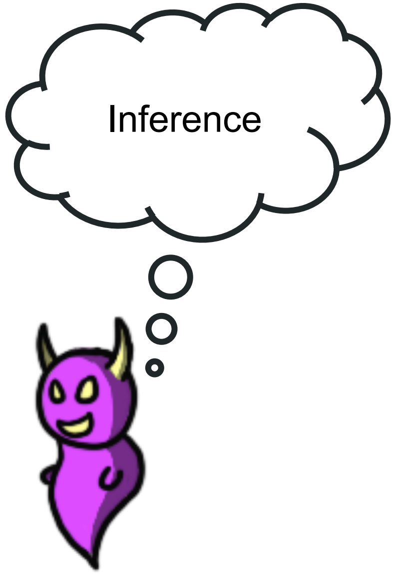
- We will always remove identifiers and focus on confidential attributes.

System Model

- Each user contributes to a row in a database
- A data curator releases a sanitized version of the database
- The adversary/analyst sees the sanitized database

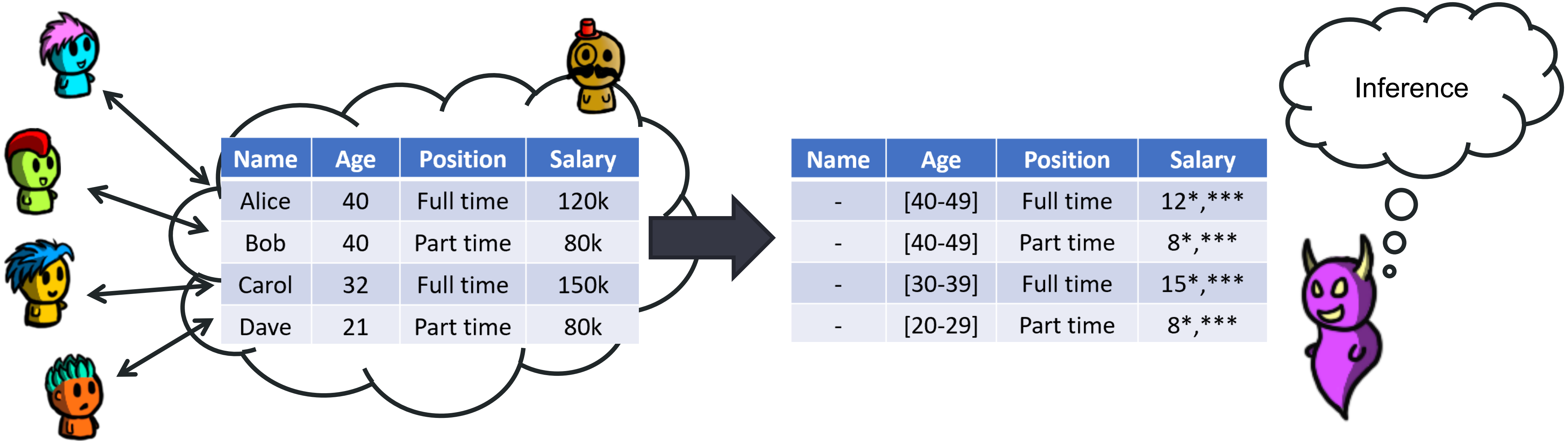
Name	Age	Position	Salary
Alice	40	Full time	120k
Bob	40	Part time	80k
Carol	32	Full time	150k
Dave	21	Part time	80k

Name	Age	Position	Salary
-	[40-49]	Full time	12*,***
-	[40-49]	Part time	8*,***
-	[30-39]	Full time	15*,***
-	[20-29]	Part time	8*,***



System Model

Q: What are the properties the sanitized database should have to preserve some level of privacy to its users?



Name	Age	Position	Salary
Alice	40	Full time	120k
Bob	40	Part time	80k
Carol	32	Full time	150k
Dave	21	Part time	80k

Name	Age	Position	Salary
-	[40-49]	Full time	12*,***
-	[40-49]	Part time	8*,***
-	[30-39]	Full time	15*,***
-	[20-29]	Part time	8*,***

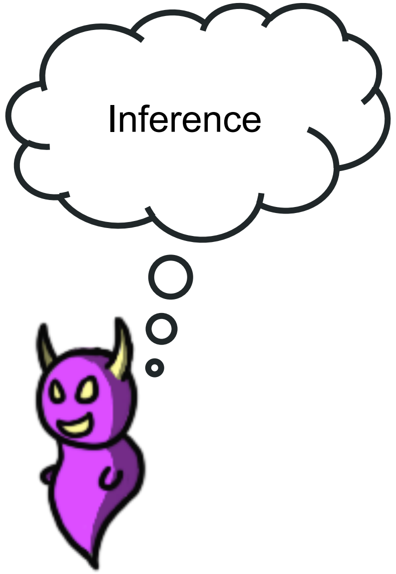
System Model

Q: What are the properties the sanitized database should have to preserve some level of privacy to its users?

A:
- k-anonymity
- ℓ-diversity
- t-closeness

Name	Age	Position	Salary
Alice	40	Full time	120k
Bob	40	Part time	80k
Carol	32	Full time	150k
Dave	21	Part time	80k

Name	Age	Position	Salary
-	[40-49]	Full time	12*,***
-	[40-49]	Part time	8*,***
-	[30-39]	Full time	15*,***
-	[20-29]	Part time	8*,***



k-anonymity

k-anonymity

For each published record, there exists at least $k-1$ other records with the same quasi-identifiers

- To **compute** k-anonymity:
  - Group the rows with the same quasi-identifier(s).
    - These rows form an *equivalence class* or equi-class.
  - **Count:** what is the smallest size of a group? That will be the level of k-anonymity
- To **provide** k-anonymity:
  - Remove a quasi-identifier
  - Reduce the granularity of a quasi-identifier (e.g., hiding the last characters of a ZIP code)
  - Group quasi-identifiers (e.g., report age ranges instead of actual ages)

k-anonymity: example

ZIP (QI)	Party affiliation
N1CFFA	Green Party
G0ANFA	Liberal Party
N1C5YN	Green Party
N2J0HJ	Conservative Party
N1C4KH	Green Party
G0A3G4	Conservative Party
G0A3GN	Liberal Party
N2JWBV	New Democratic Party
N2JWBV	Liberal Party

⇨

ZIP	Party affiliation
N1C***	Green Party
G0A***	Liberal Party
N1C***	Green Party
N2J***	Conservative Party
N1C***	Green Party
G0A***	Conservative Party
G0A***	Liberal Party
N2J***	New Democratic Party
N2J***	Liberal Party

Q: what is the k-anonymity level?

k-anonymity: example

ZIP (QI)	Party affiliation
N1CFFA	Green Party
G0ANFA	Liberal Party
N1C5YN	Green Party
N2J0HJ	Conservative Party
N1C4KH	Green Party
G0A3G4	Conservative Party
G0A3GN	Liberal Party
N2JWBV	New Democratic Party
N2JWBV	Liberal Party

⇨

ZIP	Party affiliation
N1C***	Green Party
G0A***	Liberal Party
N1C***	Green Party
N2J***	Conservative Party
N1C***	Green Party
G0A***	Conservative Party
G0A***	Liberal Party
N2J***	New Democratic Party
N2J***	Liberal Party

Q: what is the k-anonymity level?

A: the table is 3-anonymous

k-anonymity: example (II)

ZIP (QI)	DOB (QI)	Party affiliation
N1CFF	1962-01-24	Green Party
G0ANF	1975-12-30	Liberal Party
N1C5YN	1966-10-17	Green Party
N2J0HJ	1996-08-14	Conservative Party
N1C4KH	1963-04-06	Green Party
G0A3G4	1977-07-09	Conservative Party
G0A3GN	1973-08-14	Liberal Party
N2JWBV	1990-11-02	New Democratic Party
N2JWBV	1990-01-25	Liberal Party

⇨

ZIP	DOB	Party affiliation
N1C***	196*-**-**	Green Party
G0A***	197*-**-**	Liberal Party
N1C***	196*-**-**	Green Party
N2J***	199*-**-**	Conservative Party
N1C***	196*-**-**	Green Party
G0A***	197*-**-**	Conservative Party
G0A***	197*-**-**	Liberal Party
N2J***	199*-**-**	New Democratic Party
N2J***	199*-**-**	Liberal Party

Q: what is the k-anonymity level?

k-anonymity: example (II)

ZIP (QI)	DOB (QI)	Party affiliation
N1CFF	1962-01-24	Green Party
G0ANF	1975-12-30	Liberal Party
N1C5YN	1966-10-17	Green Party
N2J0HJ	1996-08-14	Conservative Party
N1C4KH	1963-04-06	Green Party
G0A3G4	1977-07-09	Conservative Party
G0A3GN	1973-08-14	Liberal Party
N2JWBV	1990-11-02	New Democratic Party
N2JWBV	1990-01-25	Liberal Party

⇨

ZIP	DOB	Party affiliation
N1C***	196*-**-**	Green Party
G0A***	197*-**-**	Liberal Party
N1C***	196*-**-**	Green Party
N2J***	199*-**-**	Conservative Party
N1C***	196*-**-**	Green Party
G0A***	197*-**-**	Conservative Party
G0A***	197*-**-**	Liberal Party
N2J***	199*-**-**	New Democratic Party
N2J***	199*-**-**	Liberal Party

Q: what is the k-anonymity level?

A: the table is 3-anonymous

k-anonymity: practice

- Both age and gender are QI.

Age	Gender	...
23	F	
25	F	
33	F	
35	F	
27	M	
30	M	
32	M	
21	NB	
25	NB	

Q: What is the k-anonymity if…

- We hide the Age
- We hide the Gender (but not the age)
- We report the most significant digit of Age, plus the Gender
- We only report the most significant digit of Age, but not the Gender

k-anonymity: practice

- Both age and gender are QI.

Age	Gender	...
23	F	
25	F	
33	F	
35	F	
27	M	
30	M	
32	M	
21	NB	
25	NB	

Q: What is the k-anonymity if…
- We hide the Age
- We hide the Gender (but not the age)
- We report the most significant digit of Age, plus the Gender
- We only report the most significant digit of Age, but not the Gender

A: 2, 1, 1, 4

k-anonymity: practice (II)

- Both age and DOB are QI.

Gender	DOB	Party affiliation
M	1968-**-**	Green Party
F	1975-**-**	Liberal Party
O	1966-**-**	Green Party
M	1962-**-**	Green Party
M	1962-**-**	Conservative Party
O	1966-**-**	Conservative Party
F	1973-**-**	Liberal Party
F	1973-**-**	Liberal Party
O	1968-**-**	Green Party
F	1975-**-**	Liberal Party

Q: What is the k-anonymity if…
- We publish the table as shown
- We hide the least-significant digit of year
- We hide the Gender column
- We hide the least-significant digit of year and hide the Gender column

k-anonymity: practice (II)

- Both age and DOB are QI.

Gender	DOB	Party affiliation
M	1968-**-**	Green Party
F	1975-**-**	Liberal Party
O	1966-**-**	Green Party
M	1962-**-**	Green Party
M	1962-**-**	Conservative Party
O	1966-**-**	Conservative Party
F	1973-**-**	Liberal Party
F	1973-**-**	Liberal Party
O	1968-**-**	Green Party
F	1975-**-**	Liberal Party

Q: What is the k-anonymity if…

- We publish the table as shown
- We hide the least-significant digit of year
- We hide the Gender column
- We hide the least-significant digit of year and hide the Gender column

A: 1, 3, 2, 4

k-anonymity and privacy

ZIP (QI)	DOB (QI)	Party affiliation
N1C***	196*-**-**	Green Party
N1C***	196*-**-**	Green Party
N1C***	196*-**-**	Green Party
G0A***	197*-**-**	Liberal Party
G0A***	197*-**-**	Liberal Party
G0A***	197*-**-**	Conservative Party
N2J***	199*-**-**	Conservative Party
N2J***	199*-**-**	New Democratic Party
N2J***	199*-**-**	Liberal Party

- This table is 3-anonymous.

Q: This provides some resistance against linking attacks, why?

k-anonymity and privacy

ZIP (QI)	DOB (QI)	Party affiliation
N1C***	196*-**-**	Green Party
N1C***	196*-**-**	Green Party
N1C***	196*-**-**	Green Party
G0A***	197*-**-**	Liberal Party
G0A***	197*-**-**	Liberal Party
G0A***	197*-**-**	Conservative Party
N2J***	199*-**-**	Conservative Party
N2J***	199*-**-**	New Democratic Party
N2J***	199*-**-**	Liberal Party

- This table is 3-anonymous.

Q: Is k-anonymity enough? Can you see any issues with it?

k-anonymity and privacy

ZIP (QI)	**DOB (QI)**	**Party affiliation**
N1C***	196*-**-**	Green Party
N1C***	196*-**-**	Green Party
N1C***	196*-**-**	Green Party
G0A***	197*-**-**	Liberal Party
G0A***	197*-**-**	Liberal Party
G0A***	197*-**-**	Conservative Party
N2J***	199*-**-**	Conservative Party
N2J***	199*-**-**	New Democratic Party
N2J***	199*-**-**	Liberal Party

- This table is 3-anonymous.

Q: Is k-anonymity enough? Can you see any issues with it?

Attack 1: if you know Alice has ZIP code N1C***, what can you learn from her?

Attack 2: if you know Bob has ZIP code G0A*** and does not like Liberal Party, what can you learn from him?

ℓ-diversity

ℓ-diversity

For each quasi-identifier value, there should be at least ℓ distinct values of the sensitive attributes

- To **compute** ℓ-diversity:
  - Group the rows by quasi-identifiers into equi-classes.
  - For each equi-class, compute how many distinct sensitive values there are
  - The equi-class with the smallest number of distinct sensitive values is the level of ℓ-diversity.
- To **provide** ℓ-diversity:
  - Similar to k-anonymity: try to make the equi-classes as large as possible, while making sure there is enough variety of sensitive attributes per class.

ℓ-diversity: example

Gender	DOB	Party affiliation
M	196*-**-**	Green Party
M	196*-**-**	Liberal Party
M	196*-**-**	Conservative Party
O	196*-**-**	Green Party
O	196*-**-**	Green Party
O	196*-**-**	Conservative Party
F	197*-**-**	Liberal Party
F	197*-**-**	Green Party
F	197*-**-**	Conservative Party
F	197*-**-**	Liberal Party

- Gender and DOB are QI, Party affiliation is the sensitive attribute.

Q: what is the level of ℓ-diversity?

ℓ-diversity: example

Gender	DOB	Party affiliation
M	196*-**-**	Green Party
M	196*-**-**	Liberal Party
M	196*-**-**	Conservative Party
O	196*-**-**	Green Party
O	196*-**-**	Green Party
O	196*-**-**	Conservative Party
F	197*-**-**	Liberal Party
F	197*-**-**	Green Party
F	197*-**-**	Conservative Party
F	197*-**-**	Liberal Party

- Gender and DOB are QI, Party affiliation is the sensitive attribute.

Q: what is the level of ℓ-diversity?

A: the table is 2-diversified

ℓ-diversity and privacy

ZIP	DOB	Salary
N3P***	199*-**-**	20K
N3P***	199*-**-**	15K
N3P***	199*-**-**	25K
H1A***	196*-**-**	100K
H1A***	196*-**-**	90K
H1A***	196*-**-**	120K
S4N***	197*-**-**	50K
S4N***	197*-**-**	60K
S4N***	197*-**-**	65K

Q: what is the level of k-anonymity and ℓ-diversity?

ℓ-diversity and privacy

ZIP	DOB	Salary
N3P***	199*-**-**	20K
N3P***	199*-**-**	15K
N3P***	199*-**-**	25K
H1A***	196*-**-**	100K
H1A***	196*-**-**	90K
H1A***	196*-**-**	120K
S4N***	197*-**-**	50K
S4N***	197*-**-**	60K
S4N***	197*-**-**	65K

Q: what is the level of k-anonymity and ℓ-diversity?

A: 3 and 3

Q: why does this provide privacy?

ℓ-diversity and privacy

ZIP	DOB	Salary
N3P***	199*-**-**	20K
N3P***	199*-**-**	15K
N3P***	199*-**-**	25K
H1A***	196*-**-**	100K
H1A***	196*-**-**	90K
H1A***	196*-**-**	120K
S4N***	197*-**-**	50K
S4N***	197*-**-**	60K
S4N***	197*-**-**	65K

Q: what is the level of k-anonymity and ℓ-diversity?

A: 3 and 3

Q: why does this provide privacy?

A: it alleviates the problem of k-anonymity when all values are the same.

Q: is this good enough? Do you see any issue?

ℓ-diversity and privacy

ZIP	DOB	Salary	Disease
N3P***	199*-**-**	20K	gastric ulcer
N3P***	199*-**-**	15K	gastritis
N3P***	199*-**-**	25K	stomach cancer
H1A***	196*-**-**	100K	heart attack
H1A***	196*-**-**	90K	flu
H1A***	196*-**-**	120K	bronchitis
S4N***	197*-**-**	50K	COVID
S4N***	197*-**-**	60K	kidney stone
S4N***	197*-**-**	65K	pneumonia

Q: is this good enough? Do you see any issue?

Q: if you know Charles, who earns a low salary, is in this table: what else did you learn?

ℓ-diversity and privacy

ZIP	DOB	Salary	Disease
N3P***	199*-**-**	20K	gastric ulcer
N3P***	199*-**-**	15K	gastritis
N3P***	199*-**-**	25K	stomach cancer
H1A***	196*-**-**	100K	heart attack
H1A***	196*-**-**	90K	flu
H1A***	196*-**-**	120K	bronchitis
S4N***	197*-**-**	50K	COVID
S4N***	197*-**-**	60K	kidney stone
S4N***	197*-**-**	65K	pneumonia

Q: is this good enough? Do you see any issue?

Q: if you know Charles, who earns a low salary, is in this table: what else did you learn?

A: Charles has a stomach disease (Similarity attack)

ℓ-diversity and privacy

ZIP	DOB	Virus X Test
N3P***	199*-**-**	Positive
N3P***	199*-**-**	Positive
N3P***	199*-**-**	Positive
… 45 more positive cases …		
N3P***	199*-**-**	Negative
H1A***	196*-**-**	Negative
H1A***	196*-**-**	Negative
H1A***	196*-**-**	Negative
… 945 more negative cases …		
H1A***	196*-**-**	Positive

Q: if you know David, who is in his 20s, is in this table: what else did you learn?

ℓ-diversity and privacy

ZIP	DOB	Virus X Test
N3P***	199*-**-**	Positive
N3P***	199*-**-**	Positive
N3P***	199*-**-**	Positive
… 45 more positive cases …		
N3P***	199*-**-**	Negative
H1A***	196*-**-**	Negative
H1A***	196*-**-**	Negative
H1A***	196*-**-**	Negative
… 945 more negative cases …		
H1A***	196*-**-**	Positive

Q: if you know David, who is in his 20s, is in this table: what else did you learn?

A: David probably has the virus (Skewness attack)

What went wrong?

ZIP	DOB	Virus X Test
N3P***	199*-**-**	Positive
N3P***	199*-**-**	Positive
N3P***	199*-**-**	Positive
... 45 more positive cases ...		
N3P***	199*-**-**	Negative
H1A***	196*-**-**	Negative
H1A***	196*-**-**	Negative
H1A***	196*-**-**	Negative
... 945 more negative cases ...		
H1A***	196*-**-**	Positive

- The data in each equi-class is unexpectedly skewed.
- This means that learning the equi-class of a person can leak a lot of statistical information about the sensitive attributes of that person.

t-closeness

t-closeness
The distribution of sensitive values in each equi-class is no further than a threshold t from the overall distribution of the sensitive values in the whole table

- To **compute** t-closeness:
  - Organize rows by equi-class
  - Compute the distribution of sensitive attributes per equi-class and for the whole table.
  - Compute the maximum difference between a class distribution and the whole table's distribution on a sensitive value. That's the value of t.
- To **provide** t-closeness:
  - Similar to k-anonymity: try to make the equi-classes as large as possible, while trying to maintain a uniform distribution.
  - Could add dummy records to help smooth the distribution.

t-closeness

t-closeness
The distribution of sensitive values in each equi-class is no further than a threshold t from the overall distribution of the sensitive values in the whole table

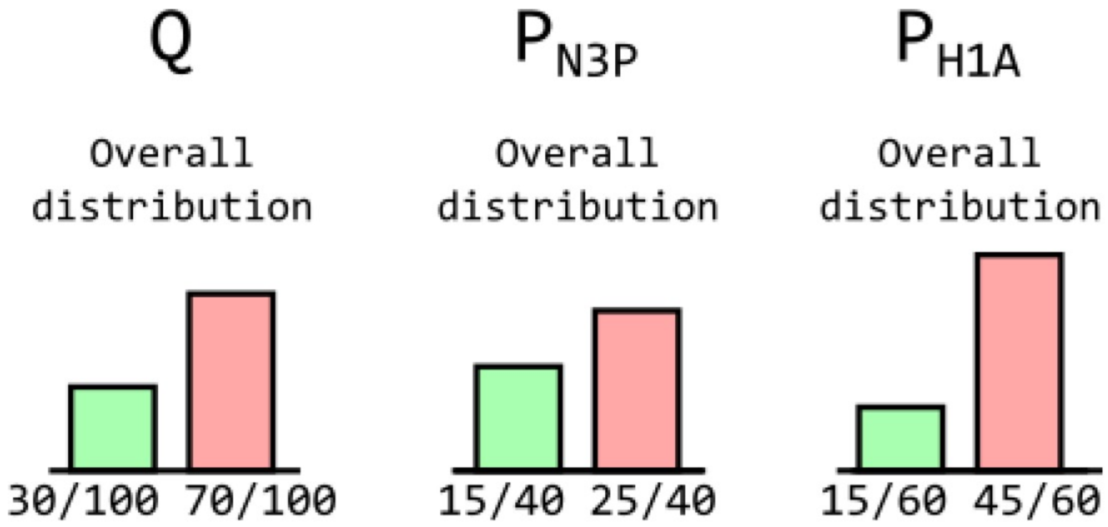
- To **compute** t-closeness we need to define a notion of distance between distributions. See the original paper that proposes t-closeness for a full description of distance notions
- We will only see one distance:

Variational distance (or EMD Categorical Distance using Equal Distance)
For two distributions over m values $P = (p_1, p_2, \dots, p_m)$ and $Q = (q_1, q_2, \dots, q_m)$:

$$D[P, Q] \doteq \frac{1}{2} \sum_{i=1}^{m} |p_i - q_i|$$

t-closeness example

ZIP (QI)	Virus (Sens)	
N3P***	Pos	x15
N3P***	Neg	x25
H1A***	Pos	x15
H1A***	Neg	x45



$$D[\mathbf{P}_{N3P}, \mathbf{Q}] = \tfrac{1}{2}\left(\left|\tfrac{15}{40} - \tfrac{30}{100}\right| + \left|\tfrac{25}{40} - \tfrac{70}{100}\right|\right) = 0.075$$

$$D[\mathbf{P}_{H1A}, \mathbf{Q}] = \tfrac{1}{2}\left(\left|\tfrac{15}{60} - \tfrac{30}{100}\right| + \left|\tfrac{45}{60} - \tfrac{70}{100}\right|\right) = 0.05$$

Variational distance:

$$D[P,Q] \doteq \frac{1}{2}\sum_{i=1}^{m} |p_i - q_i|$$

t-close with t=0.075 (the maximum of these values)

t-closeness example: more sensitive values

ZIP (QI)	Virus (Sens)	
N3P***	Pos	x5
N3P***	Neg	x22
N3P***	Inc	x3
H1A***	Pos	x12
H1A***	Neg	x47
H1A***	Inc	x1

Variational distance:

$$D[P,Q] \doteq \frac{1}{2}\sum_{i=1}^{m}|p_i - q_i|$$

Q: what is the k-anonymity, ℓ-diversity and t-closeness level of this published dataset?

A: 30-anonymous and 3-diversified.

$$D[P_{N3P},Q] = \frac{1}{2}\left(\left|\frac{5}{30}-\frac{17}{90}\right| + \left|\frac{22}{30}-\frac{69}{90}\right| + \left|\frac{3}{30}-\frac{4}{90}\right|\right) = \frac{1}{18}$$

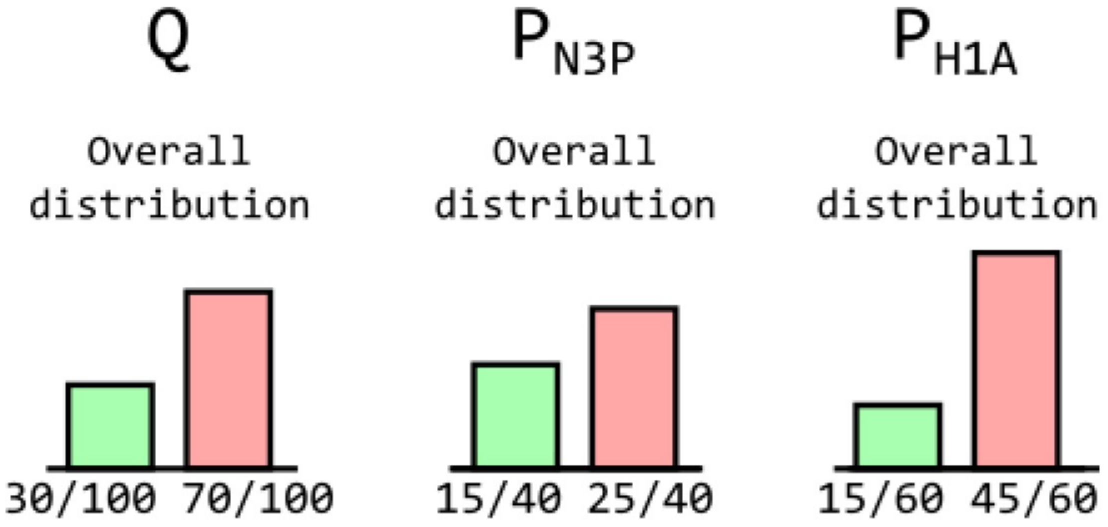
$$D[P_{H1A},Q] = \frac{1}{2}\left(\left|\frac{12}{60}-\frac{17}{90}\right| + \left|\frac{47}{60}-\frac{69}{90}\right| + \left|\frac{1}{60}-\frac{4}{90}\right|\right) = \frac{1}{36}$$

Therefore, the table is $\frac{1}{18}$-close with respect to Virus

Notes on computing t-closeness

- If you have k equi-classes, you would have to compute k distances and take the maximum of those distances as the value of t.
- If you have m distinct sensitive values, the histograms would have m bars and you would have to add m absolute value terms to compute each distance.

ZIP (QI)	Virus (Sens)	
N3P***	Pos	x15
N3P***	Neg	x25
H1A***	Pos	x15
H1A***	Neg	x45



Notes on computing t-closeness

- If you have more than one sensitive attribute (column), you can compute the t-closeness for each sensitive attribute independently (e.g., a table can be t_1-close with respect to Salary and t_2-close with respect to Virus).
- Check the original paper by Li et al. for other distance metrics and more examples.

Limitations

- t-closeness is overall a reasonable syntactic notion of privacy. It prevents the attacks that we have seen. However:

1. These privacy notions require a clear distinction between quasi-identifiers and sensitive values, which is not always possible (and is subjective)
2. Expensive to compute:
   - Computing the optimal k-anonymous dataset is NP-hard
3. These notions of privacy do not provide guarantees against an adversary with (arbitrary) background knowledge

Limitations Example

Hospital A

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<30	*	AIDS
2	130**	<30	*	Heart Disease
3	130**	<30	*	Viral Infection
4	130**	<30	*	Viral Infection
5	130**	≥40	*	Cancer
6	130**	≥40	*	Heart Disease
7	130**	≥40	*	Viral Infection
8	130**	≥40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Hospital B

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<35	*	AIDS
2	130**	<35	*	Tuberculosis
3	130**	<35	*	Flu
4	130**	<35	*	Tuberculosis
5	130**	<35	*	Cancer
6	130**	<35	*	Cancer
7	130**	≥35	*	Cancer
8	130**	≥35	*	Cancer
9	130**	≥35	*	Cancer
10	130**	≥35	*	Tuberculosis
11	130**	≥35	*	Viral Infection
12	130**	≥35	*	Viral Infection

Q: We know that Dave just had his 35^{th} birthday! He told us on his way to the hospital A. What did we learn?

Q: We know a 28 year old visited hospitals A and B. What can we infer?

Source: Ganta et al. 2008 Composition attacks and auxiliary information in data privacy

Limitations Example

Hospital A

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<30	*	AIDS
2	130**	<30	*	Heart Disease
3	130**	<30	*	Viral Infection
4	130**	<30	*	Viral Infection
5	130**	≥ 40	*	Cancer
6	130**	≥ 40	*	Heart Disease
7	130**	≥ 40	*	Viral Infection
8	130**	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Hospital B

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<35	*	AIDS
2	130**	<35	*	Tuberculosis
3	130**	<35	*	Flu
4	130**	<35	*	Tuberculosis
5	130**	<35	*	Cancer
6	130**	<35	*	Cancer
7	130**	≥ 35	*	Cancer
8	130**	≥ 35	*	Cancer
9	130**	≥ 35	*	Cancer
10	130**	≥ 35	*	Tuberculosis
11	130**	≥ 35	*	Viral Infection
12	130**	≥ 35	*	Viral Infection

Q: We know that Dave just had his 35^{th} birthday! He told us on his way to the hospital A. What did we learn?

A: Dave has Cancer

Q: We know a 28 year old visited hospitals A and B. What can we infer?

A: They likely have AIDS

Source: Ganta et al. 2008 Composition attacks and auxiliary information in data privacy

Limitations

- We need a privacy notion that is adversary-agnostic… a ***semantic*** notion of privacy, that only depends on the mechanism!
    - In the next lecture, we will see Differential Privacy (DP)