CS 458 / 658: Computer Security and Privacy

Module 6 - Data Security and Privacy

Part 2 - Attacks and defenses on data inference

Meng Xu *(University of Waterloo)*

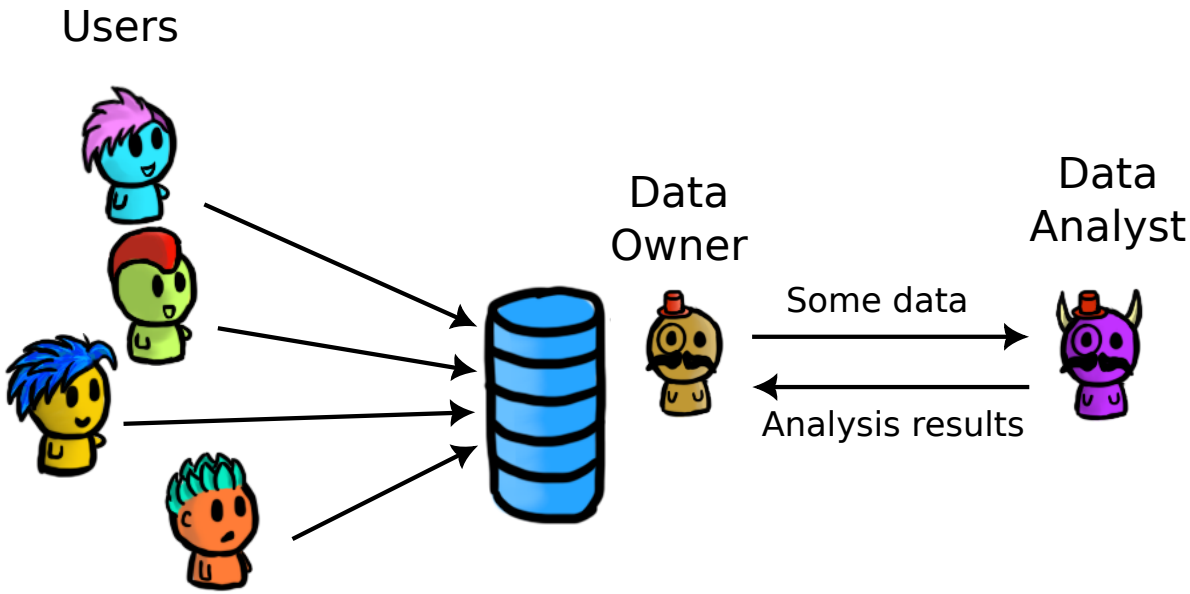
Winter 2023

Outline

System model

Consider an abstract scenario where:

- Users provide their data to a data pool
- The administrator of this data pool shares a slice of data in the pool with a data analyst.

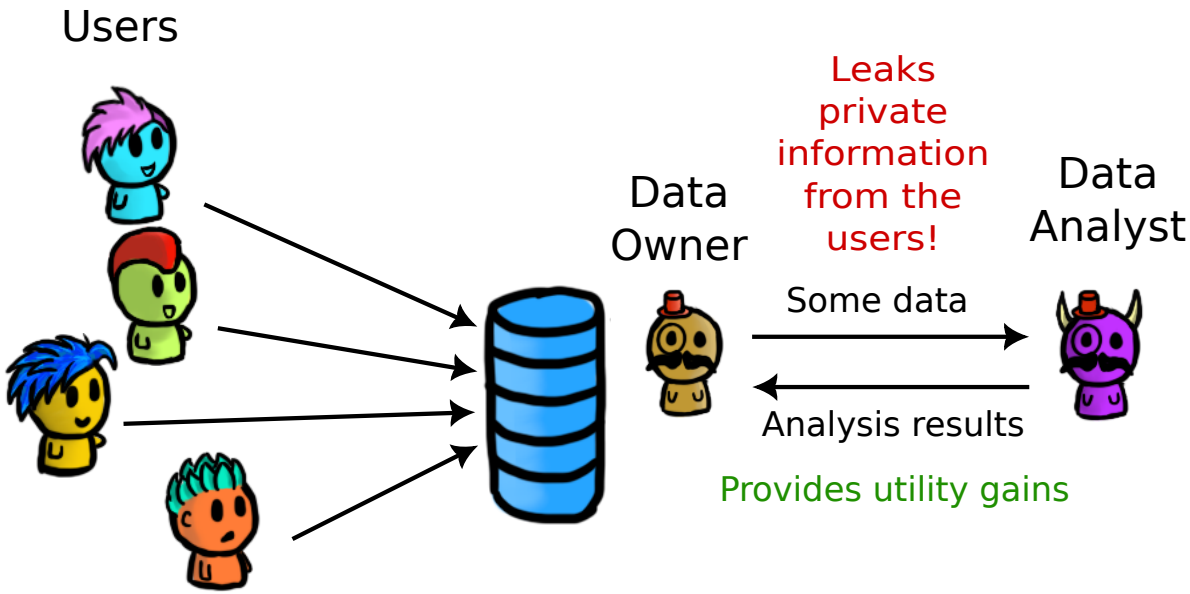


There are variations of this model, of course... (e.g., maybe the data owner/collector is a service provider that does the analysis itself)

System model

Consider an abstract scenario where:

- Users provide their data to a data pool
- The administrator of this data pool shares a slice of data in the pool with a data analyst.

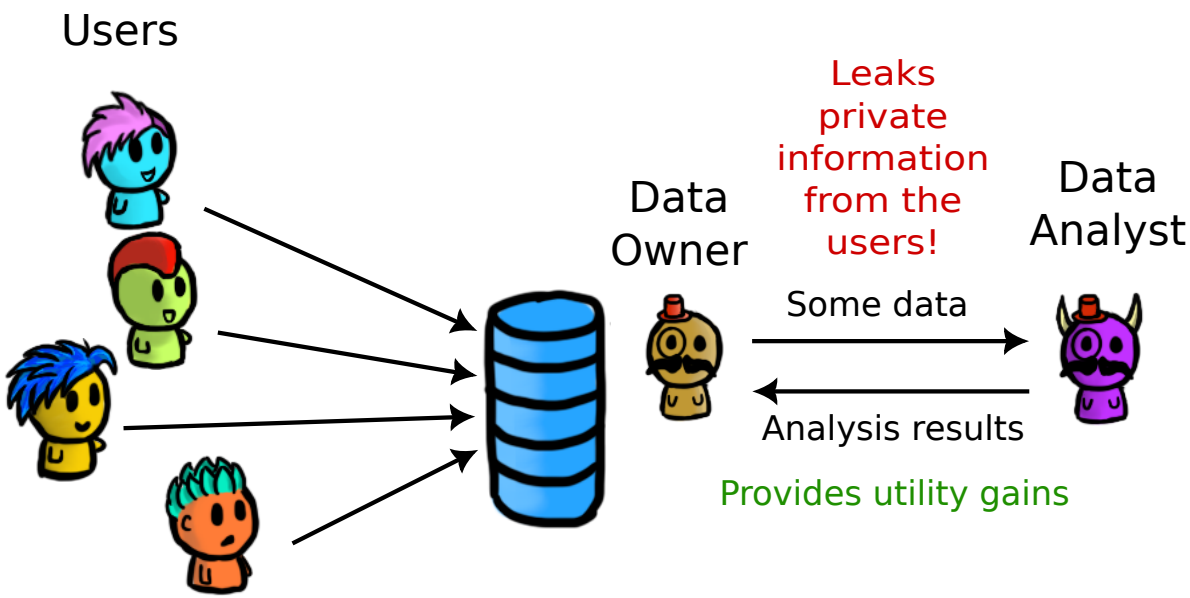


There are variations of this model, of course... (e.g., maybe the data owner/collector is a service provider that does the analysis itself)

This has privacy and utility implications!

System model: examples

Q: Any concrete examples that fit this model?

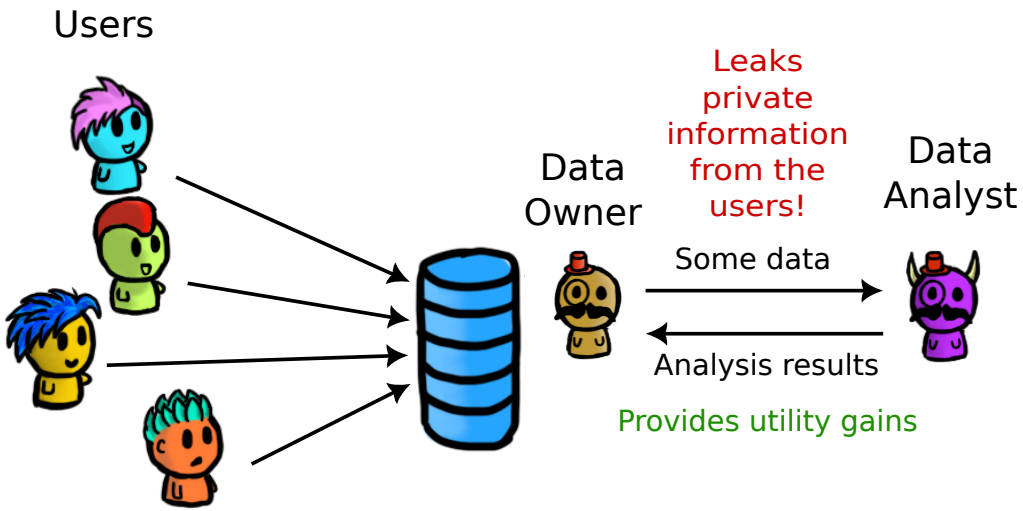


System model: examples

Q: Any concrete examples that fit this model?

Scenario	**Privacy risks**	**Utility gain**
Social media	Pictures, posts, etc.	We use social media apps for free
Virtual assistants	They hear what we say	They help us; also the recordings help improve them
Census	Personal info in census	Helps in determining how to allocate resources

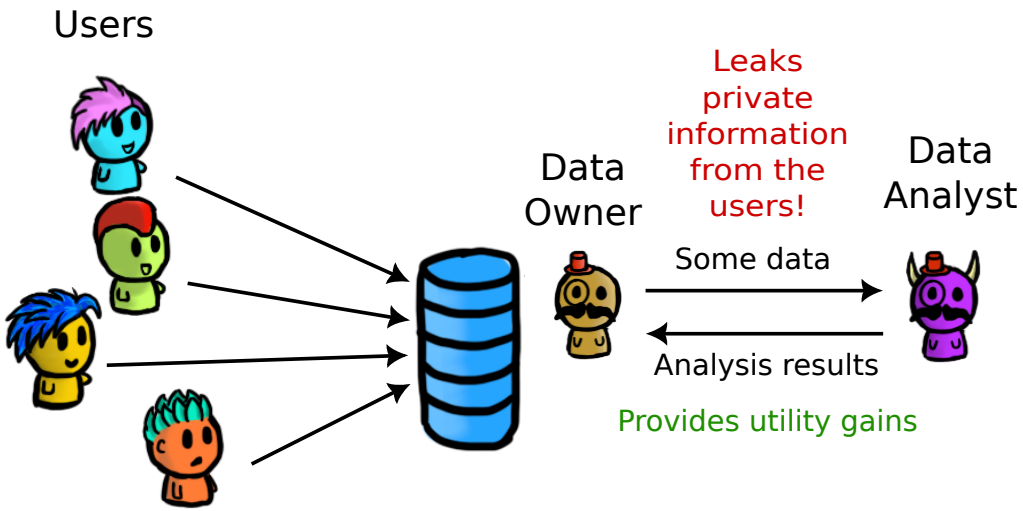
Privacy and utility



Utility can refer to benefits for both the users and the data owner/service provider.

Privacy is important for the users, since it's their data and their fundamental right to privacy.

Privacy and utility

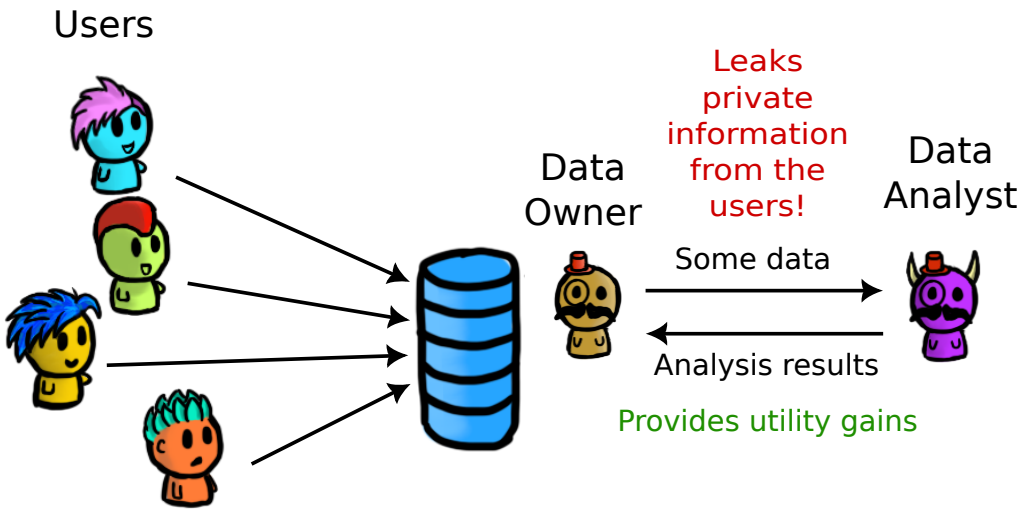


Utility can refer to benefits for both the users and the data owner/service provider.

Privacy is important for the users, since it's their data and their fundamental right to privacy.

Q: Why is privacy also important for providers?

Privacy and utility



Utility can refer to benefits for both the users and the data owner/service provider.

Privacy is important for the users, since it's their data and their fundamental right to privacy.

Q: Why is privacy also important for providers?

A: Mostly for policy compliance (e.g., GDPR)

Measuring privacy and utility

Choosing metrics for privacy and utility is not an easy task. There is no cure-all privacy metric that works for all scenarios. Same for utility (*or maybe not if you ask an economist*).

Measuring privacy and utility

Choosing metrics for privacy and utility is not an easy task. There is no cure-all privacy metric that works for all scenarios. Same for utility (*or maybe not if you ask an economist*).

We will see some **syntactic** and **semantic** notions of privacy.

- Syntactic notions: refer to some properties that the revealed/published data must follow. We will see
  - k-anonymity
  - ℓ-diversity
  - t-closeness

Measuring privacy and utility

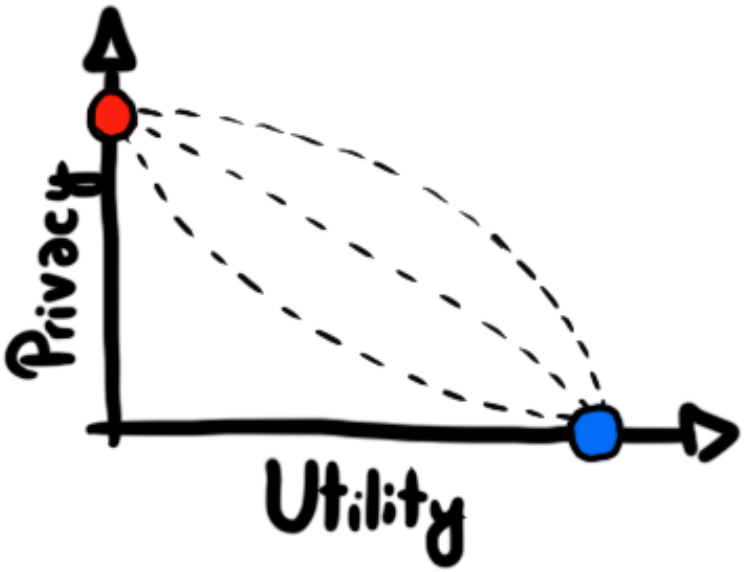
Choosing metrics for privacy and utility is not an easy task. There is no cure-all privacy metric that works for all scenarios. Same for utility (*or maybe not if you ask an economist*).

We will see some **syntactic** and **semantic** notions of privacy.

- Syntactic notions: refer to some properties that the revealed/published data must follow. We will see
  - k-anonymity
  - ℓ-diversity
  - t-closeness
- Semantic notions: refer to some properties that the data release mechanism must follow (independently of the data that is actually published!). The most popular one, which is becoming the *gold standard* for privacy, is differential privacy.

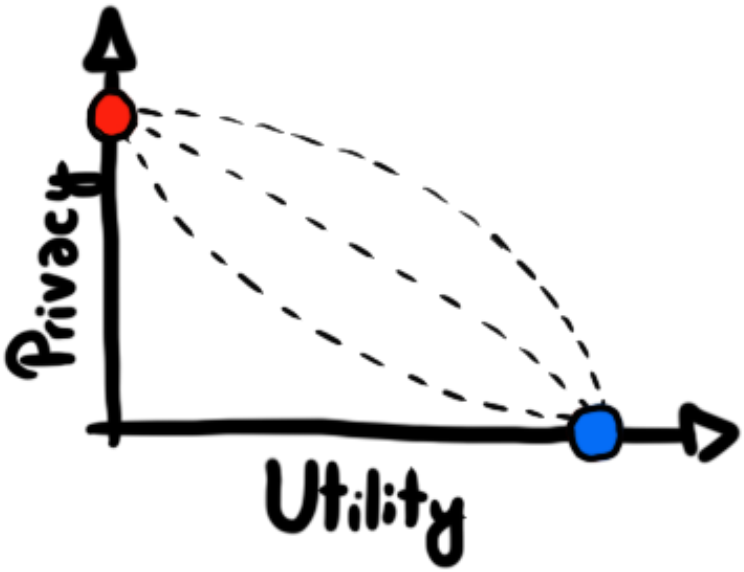
A conflict of privacy and utility

Regardless of how we quantify privacy and utility, they (often) go against each other:



A conflict of privacy and utility

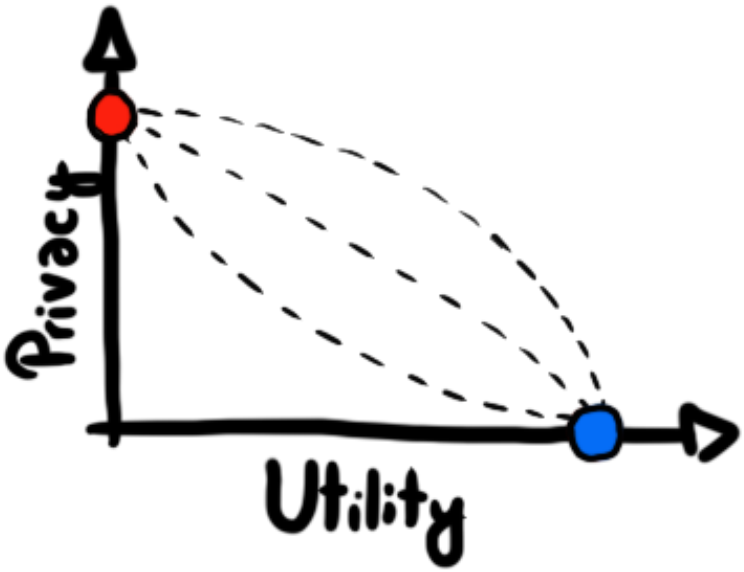
Regardless of how we quantify privacy and utility, they (often) go against each other:



Q: What's an easy approach to be in the red point? blue point?

A conflict of privacy and utility

Regardless of how we quantify privacy and utility, they (often) go against each other:



Finding data release mechanisms to be somewhere in between and enjoy a good privacy-utility trade-off is hard!

Q: What's an easy approach to be in the red point? blue point?

A:

Red point: do not provide/release/publish any data.

Blue point: release all data without protecting it.

Outline

Attacks that use SQL queries: setup

Consider a setting where we have a large relational database (e.g., a table) with some sensitive attributes.

Attacks that use SQL queries: setup

Consider a setting where we have a large relational database (e.g., a table) with some sensitive attributes.

- Utility — we want to allow certain SQL queries, as data analysts want to learn interesting properties of the data.
  - e.g., get the average salary of everyone in this company

Attacks that use SQL queries: setup

Consider a setting where we have a large relational database (e.g., a table) with some sensitive attributes.

- Utility — we want to allow certain SQL queries, as data analysts want to learn interesting properties of the data.
  - e.g., get the average salary of everyone in this company
- Privacy — We also want to protect the privacy of the users whose data is in the database.
  - e.g., without revealing each individual's salary

Attacks that use SQL queries: setup

Consider a setting where we have a large relational database (e.g., a table) with some sensitive attributes.

- Utility — we want to allow certain SQL queries, as data analysts want to learn interesting properties of the data.
  - e.g., get the average salary of everyone in this company
- Privacy — We also want to protect the privacy of the users whose data is in the database.
  - e.g., without revealing each individual's salary

Q: Can we give the permission of arbitrary read to the data analyst?

A compromise?

Now, what about a compromise solution?

- You've forbidden to issue queries that fetch a particular attribute
  - e.g., `SELECT Salary FROM Employee ...`
- but using aggregates are allowed
  - e.g., `SELECT AVG(Salary) FROM Employee ...`
  - e.g., `SELECT SUM(Salary) FROM Employee ...`
  - e.g., `SELECT COUNT(Salary) FROM Employee ...`

A compromise?

Now, what about a compromise solution?

- You've forbidden to issue queries that fetch a particular attribute
  - e.g., `SELECT Salary FROM Employee ...`
- but using aggregates are allowed
  - e.g., `SELECT AVG(Salary) FROM Employee ...`
  - e.g., `SELECT SUM(Salary) FROM Employee ...`
  - e.g., `SELECT COUNT(Salary) FROM Employee ...`

Aggregate queries that we will use

```
SELECT SUM(<Attribute>) FROM <Table> WHERE <Condition>
SELECT COUNT(*) FROM <Table> WHERE <Condition>
```

Data inference

Data inference problem: Data analysts could infer sensitive data, through output of allowed aggregate queries.

Data inference

Data inference problem: Data analysts could infer sensitive data, through output of allowed aggregate queries.

Inference does not have to be a full and accurate recovery of the sensitive data.

- e.g., the employee's salary is \$12,345.67

Instead, even a partial revealing of the data is considered as a successful inference and hence a privacy leak.

- e.g., the salary is within the range of \$10,000 and \$20,000

Data inference

Data inference problem: Data analysts could infer sensitive data, through output of allowed aggregate queries.

Inference does not have to be a full and accurate recovery of the sensitive data.

- e.g., the employee's salary is \$12,345.67

Instead, even a partial revealing of the data is considered as a successful inference and hence a privacy leak.

- e.g., the salary is within the range of \$10,000 and \$20,000

Our goal is to minimize (unintentional) leaks of sensitive data to the data analysts through the allowed queries.

Inference attack: single query

One single query that directly outputs the sensitive data

Direct attack

```
SELECT SUM(Salary) FROM Employee
  WHERE Name = "Alice"
  AND (Gender = "M" OR Gender = "F" OR Gender = "U");
```

Inference attack: single query

One single query that directly outputs the sensitive data

Direct attack

```
SELECT SUM(Salary) FROM Employee
  WHERE Name = "Alice"
  AND (Gender = "M" OR Gender = "F" OR Gender = "U");
```

Countermeasure: If the SELECT clause output includes less than k results, then drop the query. k is usually application specific.

Inference attack: multiple queries

Now, with this k value as a countermeasure, what can we do?

Inference attack: multiple queries

Now, with this k value as a countermeasure, what can we do?

Name (PK)	**Age**	**Zip**	**Salary**
Alice	32	N2L 0G7	55 000 CAD
Bob	34	N2L 3E4	65 000 CAD
Carol	26	N2L 0E1	35 000 CAD
Dave	24	N2L 2W4	40 000 CAD
...			

Table: Employee (example only)

Q: How will you infer Alice's salary in this case?

Inference attack: multiple queries

We can use set theory to dictate what queries to send, such that when their outputs are combined, the sensitive value is revealed.

Indirect attack

Q_1: `SELECT SUM(Salary) FROM Employee;` (outputs s)
Q_2: `SELECT SUM(Salary) FROM Employee WHERE Name != "Alice";` (outputs r)

$s - r$ reveals the secret salary.

Inference attack: multiple queries

We can use set theory to dictate what queries to send, such that when their outputs are combined, the sensitive value is revealed.

Indirect attack

Q_1: `SELECT SUM(Salary) FROM Employee;` (outputs s)
Q_2: `SELECT SUM(Salary) FROM Employee WHERE Name != "Alice";` (outputs r)

$s - r$ reveals the secret salary.

Countermeasure: Suppose the database has a total of N records. If the `SELECT` clause output includes less than k results, or more than $N - k$ results (but less than N results), then drop the query. *NOTE*: a query that includes N records (i.e., all records) is OK.

Inference attack: tracker attack

How do we overcome the $k \leq |Q| \leq N - k$ countermeasure?

Inference attack: tracker attack

How do we overcome the $k \leq |Q| \leq N - k$ countermeasure?

Name (PK)	**Age**	**Zip**	**Salary**
Alice	?	?	???
⋮	⋮	⋮	
⋮	⋮	⋮	
⋮	⋮	⋮	

Assumptions:

- "Alice" is in the dataset, but you don't know anything else.
- The median age in the company is 30.

Q: How will you infer Alice's salary in this case?

Inference attack: tracker attack

Assumptions:

- "Alice" is in the dataset, but you don't know anything else.
- The median age in the company is 30.

Q: How will you infer Alice's salary in this case?
Hint: the private data can be inferred with three queries

Template

Q_1: `SELECT SUM(Salary) FROM Employee WHERE ;`
Q_2: `SELECT SUM(Salary) FROM Employee WHERE ;`
Q_3: `SELECT SUM(Salary) FROM Employee WHERE ;`

Inference attack: tracker attack

Assumptions:

- "Alice" is in the dataset, but you don't know anything else.
- The median age in the company is 30.

Q: How will you infer Alice's salary in this case?
Hint: the private data can be inferred with three queries

Template

Q_1: `SELECT SUM(Salary) FROM Employee WHERE true;`
Q_2: `SELECT SUM(Salary) FROM Employee WHERE Name = "Alice" OR Age < 30;`
Q_3: `SELECT SUM(Salary) FROM Employee WHERE Name = "Alice" OR Age >= 30;`

Inference attack: tracker attack

How do we overcome the $k \leq |Q| \leq N - k$ countermeasure?

Suppose that we find a query T that satisfies this constraint:

- e.g., `SELECT SUM(Salary) FROM Employee WHERE Age < 30;`

For genericity, we use C to represent the (`Age < 30`) constraint that makes T to include a proper number of records.
And this query T is called a tracker.

Inference attack: tracker attack

How do we overcome the $k \leq |Q| \leq N - k$ countermeasure?

Suppose that we find a query T that satisfies this constraint:

- e.g., `SELECT SUM(Salary) FROM Employee WHERE Age < 30;`

For genericity, we use C to represent the (`Age < 30`) constraint that makes T to include a proper number of records.
And this query T is called a tracker.

Tracker attack

Q_1: `SELECT SUM(Salary) FROM Employee WHERE Name = "Alice" OR` C`;`
Q_2: `SELECT SUM(Salary) FROM Employee WHERE Name = "Alice" OR NOT` C`;`
Q_3: `SELECT SUM(Salary) FROM Employee;`

$Q_1 + Q_2 - Q_3$ reveals the secret salary.

What we learned from these exercises?

Having controls on the type and shape of queries is unlikely be sufficient. We need better (and more systematic) solutions to protect data privacy.

The census reconstruction attack

All the examples shown here involve a database that interactively respond to the attacker's queries, what if I do a one-time release of aggregated data only? For example, the census data?

The census reconstruction attack

All the examples shown here involve a database that interactively respond to the attacker's queries, what if I do a one-time release of aggregated data only? For example, the census data?

Suppose that we have some statistical data about a Census block:

1. There are four people in total.
2. Two of these people have age 17.
3. Two of these people self-identify as White.
4. Two of these people self-identify as Asian.
5. The average age of people who self-identify as White is 30.
6. The average age of people who self-identify as Asian is 32.

Q: Can you guess the age of everyone in the dataset?

US Census Bureau's reconstruction attack

When we have millions of statistics with many more attributes to work with, we can convert the data into a massive system of equations and use computers to solve them. See Damien Desfontaines' blog.

TLDR: The team at the Census Bureau took statistical data from the 2010 Census, transformed it into many equations, and used Gurobi to reconstruct the raw data. The records they obtained matched 46% of the original records exactly.

Outline

Inference across multiple sources

What we have seen so far uses information in a single database only. 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.

Inference across multiple sources

What we have seen so far uses information in a single database only. 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: Why more severe?

Inference across multiple sources

What we have seen so far uses information in a single database only. 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: Why more severe?

A: Because access controls rarely apply across data sources

Obtaining data sources

Where does the adversary get external data sources?

Obtaining data sources

Where does the adversary get 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.

Data linking

Now, what can we learn from combining these datasets that we didn't learn before?

Data linking

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.

Data linking

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?

Data linking

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
  - Thelma Arnold, 62-year old widow from Lilburn, GA

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 level of nymity.

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
- These parameters were collected in order 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).

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. $\Longrightarrow$ You look for pick-up records near JFK around 6pm and see where they drop-off. $\Longrightarrow$ 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. $\Longrightarrow$ You look for pick-up records near JFK around 6pm and see where they drop-off. $\Longrightarrow$ After filter out infeasible locations, you might be able to identify the taxi that they took and deduce where they lived or visited.

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

Does hashing help with hiding identities of the drivers and taxicabs?

Anonymity failure: NYC Taxi dataset release

Does hashing help with hiding identities of the drivers and taxicabs?

Background: These two identifiers have the following structures:

- 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

Does hashing help with hiding identities of the drivers and taxicabs?

Background: These two identifiers have the following structures:

- 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

Does hashing help with hiding identities of the drivers and taxicabs?

Background: These two identifiers have the following structures:

- 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. With a small input space, a dictionary attack can be conducted efficiently.

Anonymity failure: Massachusetts Insurance Health Records

Massachusetts released "anonymized" health records:

- ZIP code
- Gender
- Date of birth
- Health information

Anonymity failure: Massachusetts Insurance Health Records

Massachusetts released "anonymized" health records:

- ZIP code
- Gender
- Date of birth
- Health information

Massachusetts' voter registration lists contains:

- ZIP code
- Gender
- Date of birth
- Name

Fun fact: 87% of U.S. population can be uniquely identified using ZIP code, gender, and date of birth!

Lessons learned

Lessons learned

- Datasets included data that was useful for research (primary data), as well as some identifiers ("quasi-identifiers").

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).

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.

Privacy vs utility trade-off

What can be done about each type of data in these data releases?

Privacy vs utility trade-off

What can be done about each type of data in these data releases?

For **quasi-identifiers**:

- Reduce granularity to *deter* linking: e.g. year instead of DOB, only first couple digits of zip code. $\Longrightarrow$ Increases anonymity set.
- Remove attribute(s) to *prevent* linking altogether: e.g. no random number in AOL dataset or no medallion/license number in NYC taxi dataset. Will reduce utility of the dataset.

Privacy vs utility trade-off

What can be done about each type of data in these data releases?

For **quasi-identifiers**:

- Reduce granularity to *deter* linking: e.g. year instead of DOB, only first couple digits of zip code. $\Longrightarrow$ Increases anonymity set.
- Remove attribute(s) to *prevent* linking altogether: e.g. no random number in AOL dataset or no medallion/license number in NYC taxi dataset. Will reduce utility of the dataset.

For **primary data**:

- Reduce granularity
- Remove sensitive attributes
- Publish aggregate statistics
- Change values slightly (add randomness)

Privacy vs utility trade-off

What can be done about each type of data in these data releases?

For **quasi-identifiers**:

- Reduce granularity to *deter* linking: e.g. year instead of DOB, only first couple digits of zip code. $\Longrightarrow$ Increases anonymity set.
- Remove attribute(s) to *prevent* linking altogether: e.g. no random number in AOL dataset or no medallion/license number in NYC taxi dataset. Will reduce utility of the dataset.

For **primary data**:

- Reduce granularity
- Remove sensitive attributes
- Publish aggregate statistics
- Change values slightly (add randomness)

Syntactic notions of privacy

Syntactic notions of privacy ensure that the released data satisfies a certain property.

Syntactic notions of privacy

Syntactic notions of privacy ensure that the released data satisfies a certain property.

The data to be protected is typically a table, and the set of attributes 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: attributes (columns) that contains privacy-sensitive information.
- Non-confidential attributes: are not considered sensitive

Syntactic notions of privacy

We are going to see three syntactic notions of privacy:

- k-anonymity
- ℓ-diversity
- t-closeness

For each syntactic notion of privacy, you will learn (and need to know):

- What it **is**
- Why it provides **privacy**
- How to **compute** it
- How to **provide** it (e.g., by publishing data in a privacy-preserving way by following certain – given – utility rules)

Outline

k-anonymity

k-**anonymity**: For each published record, there exists at least $k-1$ other records with the same quasi-identifier (where $k \geq 2$).

k-anonymity

k-**anonymity**: For each published record, there exists at least $k-1$ other records with the same quasi-identifier (where $k \geq 2$).

This can be achieved by pre-processing quasi-identifiers such as

- Removing a quasi-identifier
  - e.g., removing the gender attribute
- Reducing the granularity
  - e.g., hiding the last characters of a ZIP code or the day from a DOB
- Grouping quasi-identifiers
  - e.g., reporting age ranges, instead of actual ages

k-anonymity example, with a single quasi-identifier

A simple dataset, where the quasi-identifier is ZIP.

ZIP	**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

Q: How would you apply *k*-anonymity on this table?

k-anonymity example, with a single quasi-identifier

One possibility: we hide the last three characters of ZIP, then we publish the table:

ZIP	**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

k-anonymity example, with a single quasi-identifier

One possibility: we hide the last three characters of ZIP, then we publish the table:

ZIP	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 level of *k*-anonymity?

k-anonymity example, with a single quasi-identifier

One possibility: we hide the last three characters of ZIP, then we publish the table:

ZIP	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 level of *k*-anonymity?

A: The table is 3-anonymous

k-anonymity example, with multiple quasi-identifiers

A simple dataset table (quasi-identifiers are ZIP and DOB)

ZIP	**DOB**	**Party affiliation**
N1CFF	1962-01-24	Green Party
G0ANF	1965-12-30	Liberal Party
N1C5YN	1966-10-17	Green Party
N1C0HJ	1996-08-14	Conservative Party
N1C4KH	1963-04-06	Green Party
G0A3G4	1967-07-09	Conservative Party
G0A3GN	1963-08-14	Liberal Party
N1CWBV	1990-11-02	New Democratic Party
N1CWBV	1990-01-25	Liberal Party

k-anonymity example, with multiple quasi-identifiers

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

k-anonymity example, with multiple quasi-identifiers

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

Q: What is the k-anonymity level? (ZIP and DOB are both QI)

k-anonymity example

Q: Why does *k*-anonymity provide privacy?

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

k-anonymity example

Q: Why does k-anonymity provide privacy?

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

A: We cannot identify the actual record of a user (that provided a record) based on their quasi-identifiers. This can make it hard to guess the user's confidential attributes

k-anonymity example

Q: Is this good enough?

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

Q: If you know Alice (N1C***, 196*-**-**) is in this table, what will you learn?

Homogeneity attack

A: Alice is affliated with the Green Party

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

Homogeneity attack

A: Alice is affliated with the Green Party

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

Homogeneity attack can happen when sensitive values lack diversity. In the worst case, for a given quasi-identifier, all other data values are identical.

Background knowledge attack

Q: If you know Bob (G0A***, 196*-**-**) is in this table, and Bob does not like Liberal Party, what will you learn?

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

Background knowledge attack

Q: If you know Bob (G0A***, 196*-**-**) is in this table, and Bob does not like Liberal Party, what will you learn?

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

Background knowledge attack can help filter out infeasible values and in the worst case, narrowing down to a single value only.

Outline

ℓ-diversity

ℓ-**diversity**: For any quasi-identifier value, there should be at least ℓ distinct values of the sensitive fields

ℓ-diversity example

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

Q: What is the level of ℓ-diversity?

ℓ-diversity example

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

Q: What is the level of ℓ-diversity?

A: This table is 2-diversified

ℓ-diversity example

Q: Why does ℓ-diversity provide 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

ℓ-diversity example

Q: Why does ℓ-diversity provide 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

A: It alleviates the issues of k-anonymity that we saw above. Given someone's quasi-identifiers and access to the published database, ℓ-diversity makes it harder to guess that individual's sensitive values

ℓ-diversity example

Q: Is this good enough?

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: If you know Charles who earns a low salary is in this table, what will you learn?

Similarity attack

A: Charles has a stomach disease

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

Similarity attack

A: Charles has a stomach disease

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

Similarity attack If the sensitive values of an equi-class are different but have the same (or similar) semantic meaning, ℓ-diversity does not prevent the adversary from learning this.

Skewness attack

Q: If you know David (in his 20s) is in this table, what will you learn?

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

Skewness attack

Q: If you know David (in his 20s) is in this table, what will you learn?

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

Skewness attack: the distribution of sensitive values matters! Highly-skewed distributions leak (statistically speaking) more information about an individual's sensitive value.

Outline

What went wrong?

What went wrong?

Re-examine: If you know Charles who earns a low salary is in this table, what will you learn?

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

What went wrong?

Re-examine: If you know Charles who earns a low salary is in this table, what will you learn?

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

Finding: The concentration of stomach diseases in low-income employees is unexpected.

What went wrong?

Q: What is unexpected exactly?

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

What went wrong?

Q: What is unexpected exactly?

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

A: The "unexpected" feeling comes from the distribution of sensitive values of the whole dataset being different than the distribution of the sensitive values per class. i.e., 5% of positive rate overall vs 98% of positive rate in the first group.

Reflection

Revealing the overall distribution of the sensitive attribute in the whole dataset should be considered to have no privacy leakage.

Reflection

Revealing the overall distribution of the sensitive attribute in the whole dataset should be considered to have no privacy leakage.

- $\Longleftrightarrow$ removing all quasi-identifier attributes preserves privacy.

Reflection

Revealing the overall distribution of the sensitive attribute in the whole dataset should be considered to have no privacy leakage.

- $\iff$ removing all quasi-identifier attributes preserves privacy.
- Seems unavoidable unless willing to destroy utility.

Reflection

Revealing the overall distribution of the sensitive attribute in the whole dataset should be considered to have no privacy leakage.

- $\Longleftrightarrow$ removing all quasi-identifier attributes preserves privacy.
- Seems unavoidable unless willing to destroy utility.

However, the distribution of sensitive attribute values in each equi-class (i.e., records that share the same quasi-identifier) are not! And this is where this "unexpected feeling" comes from.

An implied definition of privacy

Privacy is measured by the information gain of an observer.

An implied definition of privacy

Privacy is measured by the information gain of an observer.

The gain is the difference between

- *prior belief*, what the observer knows *before* seeing the data, and
- *posterior belief*: what the observer knowns *after* seeing the data.

An implied definition of privacy

Privacy is measured by the information gain of an observer.

The gain is the difference between

- *prior belief*, what the observer knows *before* seeing the data, and
  - e.g., David has 5% chance of having Virus X
- *posterior belief*: what the observer knowns *after* seeing the data.
  - e.g., David has 98% chance of having Virus X

t-closeness

t-closeness: Distribution of sensitive attribute values in each equi-class should be close to that of the overall dataset. The closeness is measured by some distance calculation method and is bounded by a threshold t.

t-closeness

t-closeness: Distribution of sensitive attribute values in each equi-class should be close to that of the overall dataset. The closeness is measured by some distance calculation method and is bounded by a threshold t.

For a list of distance calculation methods, see the original paper that proposes t-closeness on ICDE'07.

Outline

Limitations

Limitations

- Requires the distinction between quasi-identifiers and sensitive attributes, which is not always possible (and very subjective)

Limitations

- Requires the distinction between quasi-identifiers and sensitive attributes, which is not always possible (and very subjective)
- It is difficult to pin down adversary's background knowledge. For example, the knowledge that a user may have even participated in the dataset helps ultimately to de-anonymize users.

Limitations

- Requires the distinction between quasi-identifiers and sensitive attributes, which is not always possible (and very subjective)
- It is difficult to pin down adversary's background knowledge. For example, the knowledge that a user may have even participated in the dataset helps ultimately to de-anonymize users.
- The privacy notions are syntactic in nature, i.e., the output satisfies the privacy properties but the adversary might be able to infer more information if the adversary knows the algorithm that produces the output.
  - Consider a simple algorithm that produces a 3-anonymized 3-diversified dataset:
    1) repeat the record 2 times and
    2) do a $+1$ and -1 on the sensitive value on each duplicated record.
  - How private is that?

Limitations

However, with these limitations said,

- k-anonymity
- ℓ-diversity
- t-closeness

is probably the best we can do IF we need to release information on an entry-by-entry basis.

But for aggregated data (one-time release or interactive queries), we have a much more powerful tool — *differential privacy*.