+

Week 09
Lecture Notes

INFS3200 Advanced Database Systems
Semester 1, 2021

Data Privacy – Part 1

Lecturer: Yanjun Zhang

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

+ Outline

- Privacy issues in dataset release (week 9)
  - Data privacy – definition, challenge
  - Privacy preserving techniques for dataset publishing
- Privacy in distributed machine learning (week 11)

+ Privacy and Data Release

NYC
Taxi & Limousine
Commission

- NYC taxi and limousine commission released 2013 trip data.
  - Start point, end point, timestamps, taxi id, fare, tip amount.
  - 173 million trips "anonymized" to remove identifying information.

+ Privacy and Data Release

- Use a simple hash to anonymize personally identifiable information (the driver's licence number) → easily reversed.
  - The data had been anonymised by hashing, a cryptographic function which is supposed to be "one-way": it's very easy to find the hash of a given piece of data, and very hard – mathematically impossible, in theory – to find the piece of data which resulted in a given hash.

    "Alex" -> a08372b70196c21a9229cf04db6b7ceb

  - Licences are all six-digit or seven-digit numbers starting with a five. That means that there are only 2m possible license numbers
  - But once the possible entries have been down to 2m different numbers, it was the matter of only minutes to determine which numbers were associated with which pieces of anonymised data.

Could yield personal details, such as drivers' addresses and income!

+ Privacy and Data Release

- What’s worse, with other publicly available data, one can link people to taxis and find out where they went
  - For example, paparazzi pictures of celebrities.

Bradley Cooper (actor)

Jessica Alba (actor)

+ Privacy and Data Release

- Not just celebrities: can find trips starting at "sensitive" locations.
  - – For example, Larry Flynt's Club

- Can find more about venue's customers.
  - "Examining one of the clusters ... only one of the five likely drop-off addresses was inhabited; a search ... revealed its resident's name. By examining other drop-offs at this address ... this gentleman also frequented ... "Rick's Cabaret" and "Flashdancers". Using websites like Spokeo and Facebook ... able to find out his ... relationship status, court records and even a profile picture!"

+ Privacy and Data Release

- In 2006, the company released the movie ratings of 500,000 anonymised customers to encourage better recommendation algorithms.
- Two researchers identified NetFlix users by comparing their "anonymous" reviews in the Netflix dataset to ones posted on the Internet Movie Database website. Revelations included identifying their political leanings and sexual orientation [Narayanan and Shmatikov. 2008].
- Eventually, the revelations led to a 2009 lawsuit from an in-the-closet lesbian mother, who sued Netflix for privacy violation.
- Lesson learned: Even if identifiers such as names and Social Security numbers have been removed, the adversary can use background knowledge and cross-correlation with other databases to re-identify individual data records.

Brokeback Mountain, 2005

+ CIA Triad of Information Security

- Confidentiality: Ensures that data or an information system is accessed by only an authorized person.
- Integrity: Integrity assures that the data or information system can be trusted. Ensures that it is edited by only authorized persons and remains in its original state when at rest.
- Availability: Data and information systems are available when required.

+ Data Privacy

The General Data Protection Regulation (GDPR) https://gdpr.eu

- Data privacy: empowering users to make their own decisions about who can process their data and for what purpose
- Data privacy is the relationship among (1) the collection & dissemination of data, (2) technology, (3) the public expectation of privacy, and (4) the legal and political issues surrounding them

+ Data Utility and Data Privacy

- The challenge of data privacy is to utilize data while protecting individual's privacy preferences & their personally identifiable information
- Sensitive information
  - Identity
    - Direct identifiers: attributes that explicitly identify individuals
    - Quasi-identifiers: attributes that in combination with others lead to identification
  - Sensitive attributes
    - Attributes that individuals are not willing to disclose, such as salary, health, religion
  - Relationship

+ An Example of Statistical Attack

- Privacy rules
  - Cannot query about individual's salary
- Attack queries:

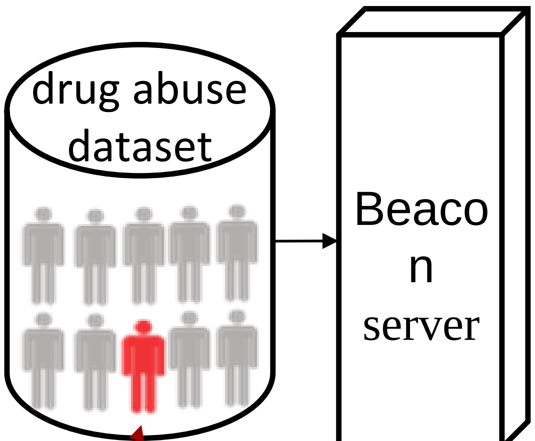
```
select  count(*)
from    staff
where   title = "Professor"

select  sum(salary)
from    staff
where   title = "Professor"
```

+ An example of real-world Statistical Attacks

Privacy rule: the drug abuse dataset is not accessible to users, but only allows them to query the allele-presence information.

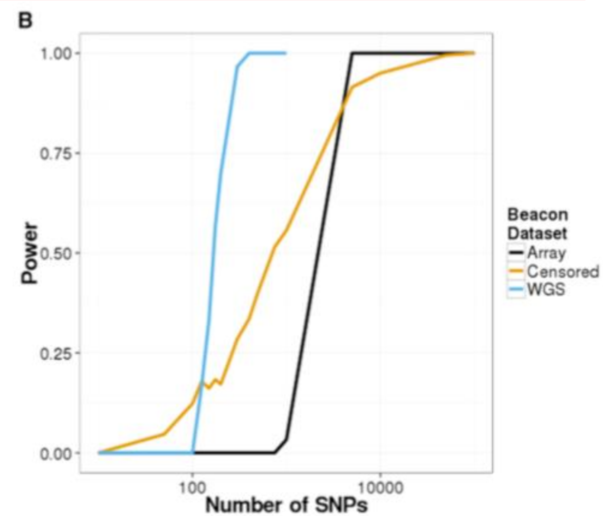
By calculating the likelihood of responses, the attacker can differentiate individuals in the beacon from those not in the beacon.



Query: Does the dataset contain any genome that has a nucleotide (e.g., A) at position 11,272 on chromosome 1?

The beacon server can answer “yes” or “no.”

Alice



Target Bob

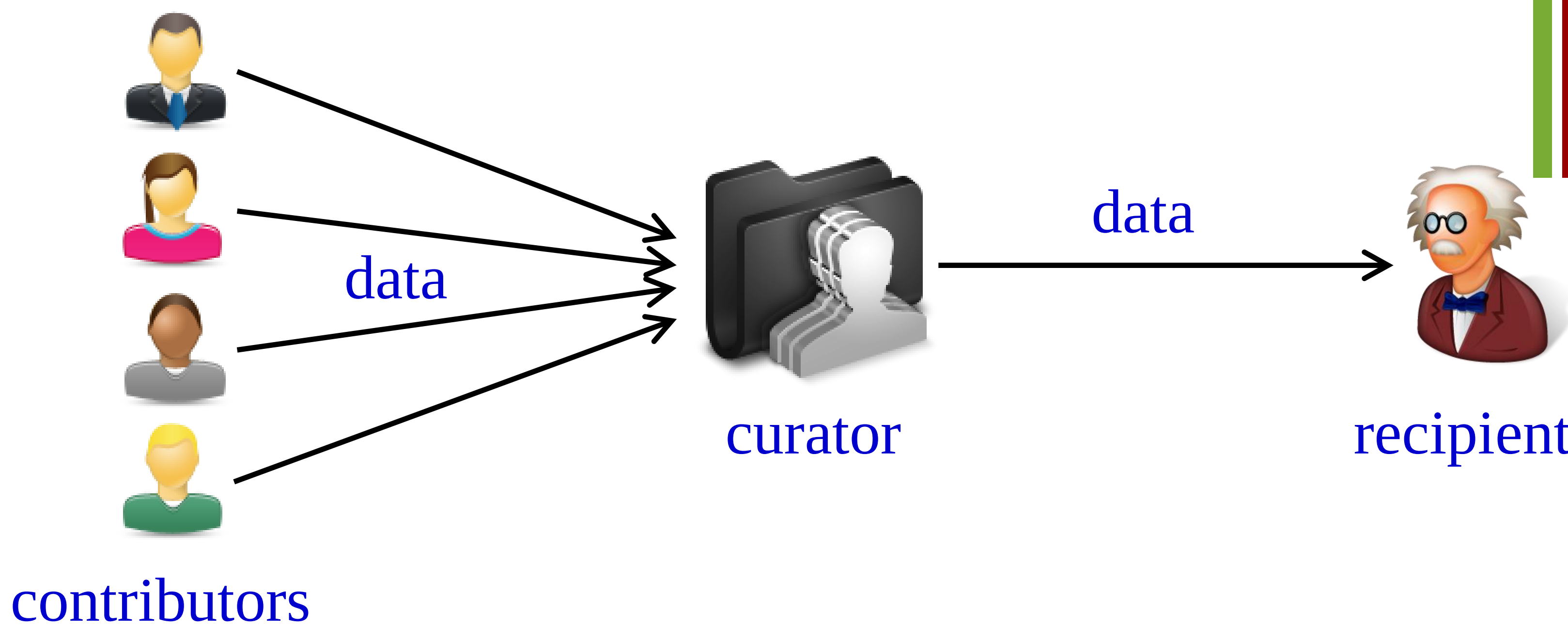
Background information:
Bob’s DNA

The Beacon Project by the Global Alliance for Genomics & Health (GA4GH) aims to simplify data sharing through a web service (“beacon”) that provides only allele-presence information.

Power of Re-identification Attacks on Beacons Constructed with Real Data: With just 250 queries, beacon membership could be detected with 95% power and a 5% false-positive rate

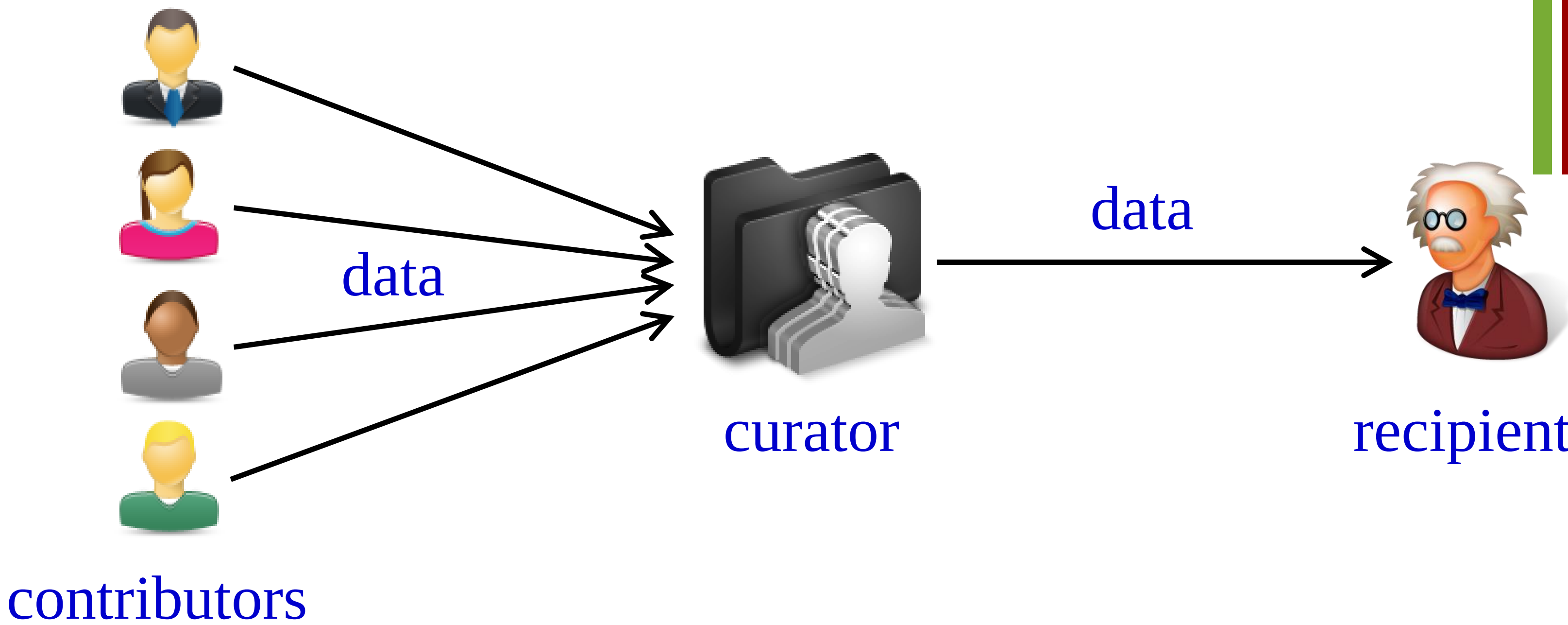
(Suyash , . et al, 2015)

+ Privacy Preserving Data Publishing



- Each contributor: provide data about herself
- Curator: collects data and releases them in a certain form
- Recipient: uses the released data for analysis

+ Privacy Preserving Data Publishing



- Objectives:
  - The privacy of the contributors are protected
  - The recipient gets useful data

+ Privacy Breach: The MGIC Case

- Curator: Massachusetts Group Insurance Commission (MGIC)
- Data released: “anonymized” medical records
- Intention: facilitate medical research

~~Name~~	Birth Date	Gender	ZIP	Disease
~~Alice~~	1960/01/01	F	10000	flu
~~Bob~~	1965/02/02	M	20000	dyspepsia
~~Cathy~~	1970/03/03	F	30000	pneumonia
~~David~~	1975/04/04	M	40000	gastritis

Medical Records

+ Privacy Breach: The MGIC Case

- Curator: Massachusetts Group Insurance Commission (MGIC)
- Data released: "anonymized" medical records
- Intention: facilitate medical research

Quasi-identifier

match

Name	Birth Date	Gender	ZIP
Alice	1960/01/01	F	10000
Bob	1965/02/02	M	20000
Cathy	1970/03/03	F	30000
David	1975/04/04	M	40000

Voter Registration List

Birth Date	Gender	ZIP	Disease
1960/01/01	F	10000	flu
1965/02/02	M	20000	dyspepsia
1970/03/03	F	30000	pneumonia
1975/04/04	M	40000	gastritis

Medical Records

+ Where Do I Get These Records?

DEC 28, 2015 @ 08:50 AM 209,785

191 Million US Voter Registration Records Leaked In Mystery Database

Thomas Fox-Brewster, FORBES STAFF
I cover crime, privacy and security in digital and physical forms. **FULL BIO**

A whitehat hacker has uncovered a database sitting on the Web containing various pieces of personal information related to 191 million American citizens registered to vote. On top of the concomitant problems of disclosing such a significant leak to that many people, no one knows who is actually responsible for the misconfiguration that left the data open to anyone.

+ Privacy Breach: The AOL Case

- Time: 2006
- Curator: American Online
- Data released: “anonymized” search log
- Intention: facilitate research on web search

- Log record: < User ID, Query, … >
- Example: < 4417749, “UQ”, … >

+ Privacy Breach: The AOL Case

- Log record: < User ID, Query, … >
- Example: < 4417749, “UQ”, … >

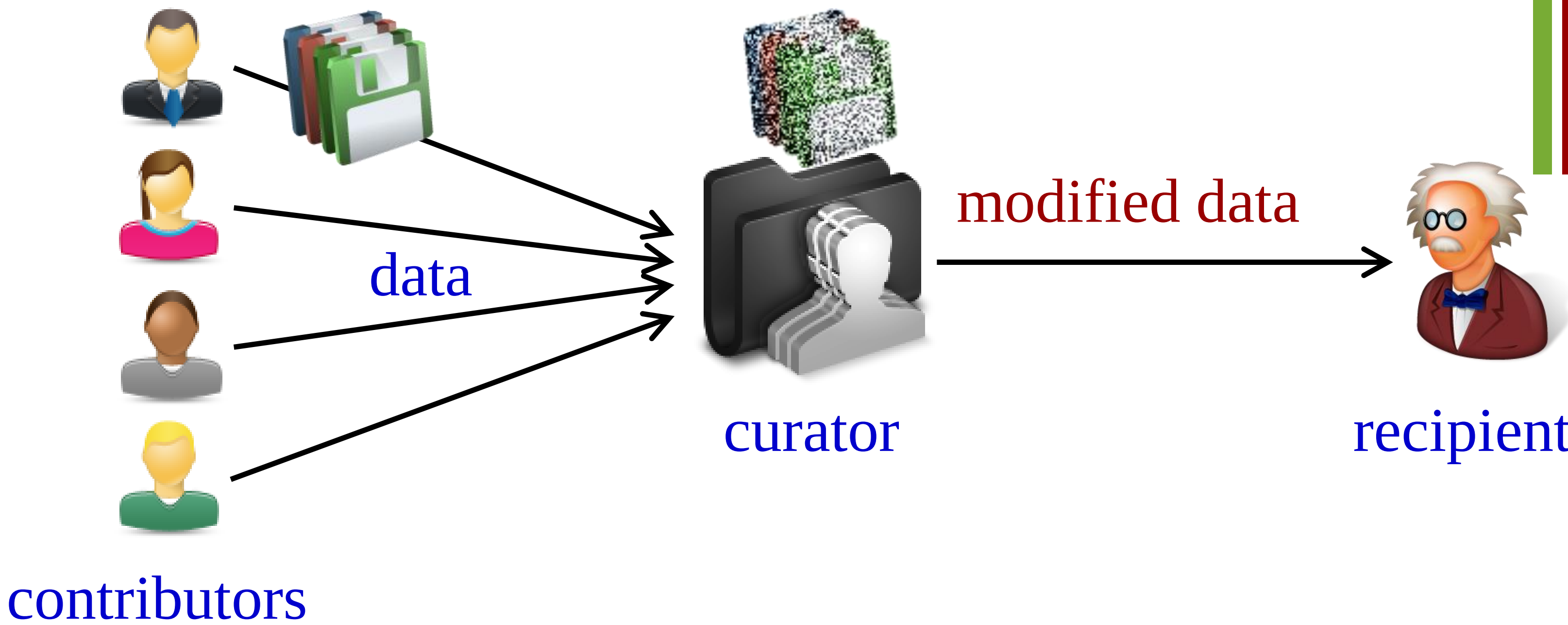
- Attacker: New York Times
- Method:
  - Find all log entries for AOL user 4417749
  - Many queries for businesses and services in Lilburn, GA (population 11K)
  - A number of queries for different persons with the last name Arnold
  - Lilburn has 14 people with the last name Arnold
  - The New York Times contacted them and found that AOL User 4417749 is Thelma Arnold

See https://en.wikipedia.org/wiki/AOL_search_data_leak

+ Lessons Learned

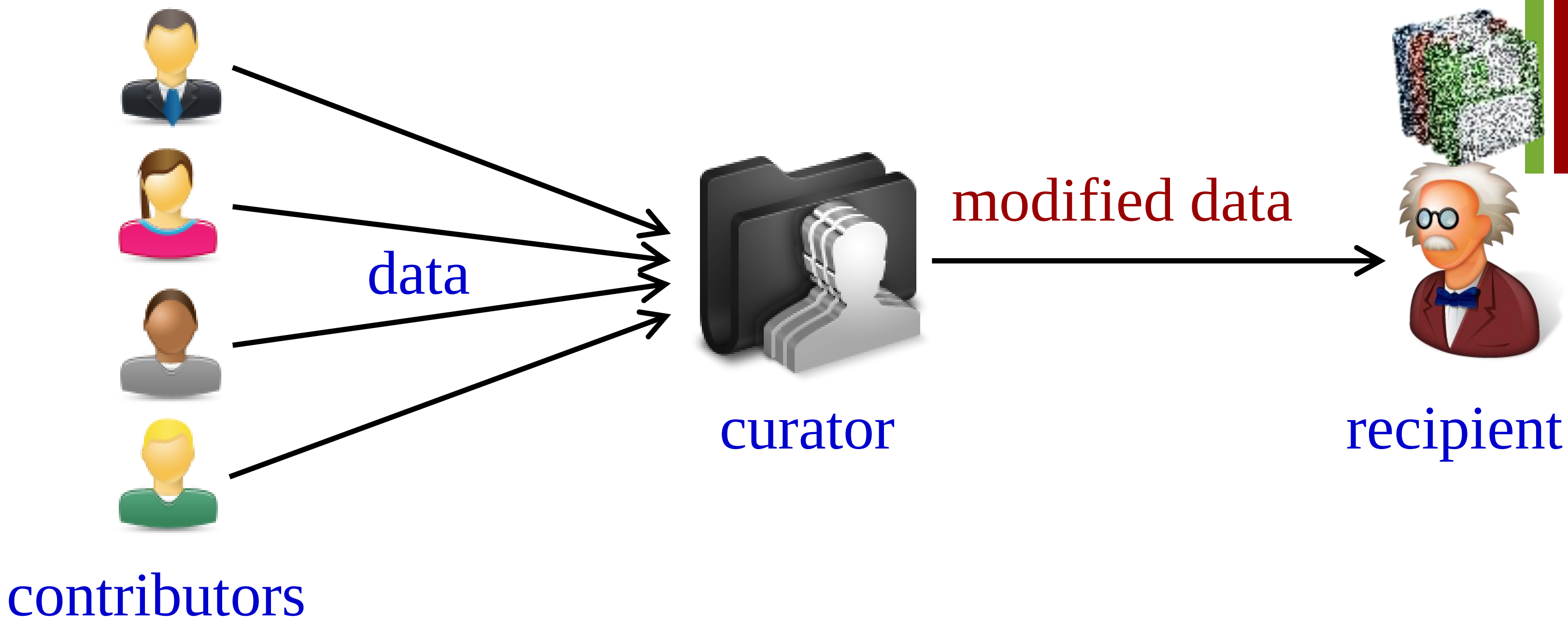
- Any information released by the data curator can potentially be exploited by the adversary
  - In the MGIC case: genders, birth dates, ZIP codes
  - In the AOL case: keywords in search queries
- Solution?
  - Do not release the exact information from the original data

+ Privacy Preserving Data Publishing



- Publish a modified version of the data, such that
  - the contributors' privacy is "adequately" protected
  - the published data is useful for its intended purpose (at least to some degree)

+ Privacy Preserving Data Publishing



- Two issues
  - privacy principle: what do we mean by “adequately” protected privacy?
  - modification method: how should we modify the data to ensure privacy while maximizing utility?

+ Existing Solutions

- Solutions before 2000
  - Mostly without a formal privacy model
  - Evaluates privacy based on empirical studies only
- This lecture will focus on solutions with formal privacy models (developed after 2000)
  - k-anonymity, l-diversity, t-closeness
  - Differential privacy

+ k-Anonymity: Example

- Suppose that we want to publish the medical records below

Name	Age	ZIP	Disease
Andy	20	10000	flu
Bob	30	20000	dyspepsia
Cathy	40	30000	pneumonia
Diane	50	40000	gastritis

+ *k*-Anonymity: Example

- Suppose that we want to publish the medical records below
- We know that
  - eliminating names is not enough
  - because an adversary may identify patients by Age and ZIP

Name	Age	ZIP
Andy	20	10000
Bob	30	20000
Cathy	40	30000
Diane	50	40000

adversary's knowledge

Age	ZIP	Disease
20	10000	flu
30	20000	dyspepsia
40	30000	pneumonia
50	40000	gastritis

medical records

+ *k*-Anonymity: Example

- *k*-anonymity [Sweeney 2002]
  - requires that each individual in a dataset is indistinguishable from *k-1* others, with respect to their **quasi-identifiers** (Age, ZIP) .
- How?
- Make Age and ZIP less specific in the medical records

Name	Age	ZIP
Andy	20	10000
Bob	30	20000
Cathy	40	30000
Diane	50	40000

adversary's knowledge

Age	ZIP	Disease
20	10000	flu
30	20000	dyspepsia
40	30000	pneumonia
50	40000	gastritis

medical records

+ *k*-Anonymity: Example

- *k*-anonymity [Sweeney 2002]
  - requires that each individual in a dataset is indistinguishable from *k-1* others, with respect to their **quasi-identifiers** (Age, ZIP) .

"generalization"

Name	Age	ZIP
Andy	20	10000
Bob	30	20000
Cathy	40	30000
Diane	50	40000

adversary's knowledge

Age	ZIP	Disease
20	10000	flu
30	20000	dyspepsia
40	30000	pneumonia
50	40000	gastritis

medical records

+ *k*-Anonymity: Example

- *k*-anonymity [Sweeney 2002]
  - requires that each individual in a dataset is indistinguishable from *k-1* others, with respect to their **quasi-identifiers** (Age, ZIP) .

Name	Age	ZIP
Andy	20	10000
Bob	30	20000
Cathy	40	30000
Diane	50	40000

adversary's knowledge

Age	ZIP	Disease
[20,30]	[10000,20000]	flu
[20,30]	[10000,20000]	dyspepsia
[40,50]	[30000,40000]	pneumonia
[40,50]	[30000,40000]	gastritis

2-anonymous table

+ *k*-Anonymity: Example

- *k*-anonymity [Sweeney 2002]
  - requires that each (Age, ZIP) combination can be matched to at least *k* patients

Name	Age	ZIP
Andy	20	10000
Bob	30	20000
Cathy	40	30000
Diane	50	40000

adversary's knowledge

Age	ZIP	Disease
[20,30]	[10000,20000]	flu
[20,30]	[10000,20000]	dyspepsia
[40,50]	[30000,40000]	pneumonia
[40,50]	[30000,40000]	gastritis

2-anonymous table

+ *k*-Anonymity: General Approach

- Identify the attributes that the adversary may know
  - Referred to as Quasi-Identifiers (QI)
- Divide tuples in the table into groups of sizes at least *k*
- Generalize the QI values of each group to make them identical

QI

group 1

group 2

Age	ZIP	Disease
20	10000	flu
30	20000	dyspepsia
40	30000	pneumonia
50	40000	gastritis

medical records

+ *k*-Anonymity: General Approach

- Identify the attributes that the adversary may know
  - Referred to as Quasi-Identifiers (QI)
- Divide tuples in the table into groups of sizes at least *k*
- Generalize the QI values of each group to make them identical

Name	Age	ZIP
Andy	20	10000
Bob	30	20000
Cathy	40	30000
Diane	50	40000

adversary's knowledge

QI

Age	ZIP	Disease
[20,30]	[10000,20000]	flu
[20,30]	[10000,20000]	dyspepsia
[40,50]	[30000,40000]	pneumonia
[40,50]	[30000,40000]	gastritis

2-anonymous table

+ *k*-Anonymity: Algorithms

- Numerous algorithms for *k*-anonymity had been proposed
- Objective: achieve *k*-anonymity with the least amount of generalization
- This line of research became obsolete
- Reason: *k*-anonymity was found to be vulnerable [Machanavajjhala et al. 2006]

Name	Age	ZIP
Andy	20	10000
Bob	30	20000
Cathy	40	30000
Diane	50	40000

adversary's knowledge

QI

Age	ZIP	Disease
[20,30]	[10000,20000]	flu
[20,30]	[10000,20000]	dyspepsia
[40,50]	[30000,40000]	pneumonia
[40,50]	[30000,40000]	gastritis

2-anonymous table

+ k-Anonymity: Vulnerability

- k-anonymity requires that each combination of quasi-identifiers (QI) is hidden in a group of size at least k
- But it says nothing about the remaining attributes
- Result: Disclosure of sensitive attributes is possible

Name	Age	ZIP
Andy	20	10000
Bob	30	20000
Cathy	40	30000
Diane	50	40000

adversary's knowledge

QI		sensitive
Age	ZIP	Disease
[20,30]	[10000,20000]	flu
[20,30]	[10000,20000]	dyspepsia
[40,50]	[30000,40000]	pneumonia
[40,50]	[30000,40000]	gastritis

2-anonymous table

+ k-Anonymity: Vulnerability

- k-anonymity requires that each combination of quasi-identifiers (QI) is hidden in a group of size at least k
- But it says nothing about the remaining attributes
- Result: Disclosure of sensitive attributes is possible

Name	Age	ZIP
Andy	20	10000
Bob	30	20000
Cathy	40	30000
Diane	50	40000

adversary's knowledge

Age (QI)	ZIP (QI)	Disease (sensitive)
[20,30]	[10000,20000]	flu
[20,30]	[10000,20000]	flu
[40,50]	[30000,40000]	pneumonia
[40,50]	[30000,40000]	gastritis

2-anonymous table

+ k-Anonymity: Vulnerability

- Intuition:
  - Hiding in a group of k is not sufficient
  - The group should have a diverse set of sensitive values

Name	Age	ZIP
Andy	20	10000
Bob	30	20000
Cathy	40	30000
Diane	50	40000

adversary's knowledge

QI sensitive

Age	ZIP	Disease
[20,30]	[10000,20000]	flu
[20,30]	[10000,20000]	flu
[40,50]	[30000,40000]	pneumonia
[40,50]	[30000,40000]	gastritis

2-anonymous table

+ l-Diversity [Machanavajjhala et al. 2006]

- Approach: (similar to k-anonymity)
  - Divide tuples into groups, and make the QI of each group identical
- Requirement: (different from k-anonymity)
  - Each group has at least l "well-represented" sensitive values
- Several definitions of "well-represented" exist
  - Simplest one: in each group, no sensitive value is associated with more than $1/l$ of the tuples

Age	ZIP	Disease
[20,30]	[10000,20000]	flu
[20,30]	[10000,20000]	dyspepsia
[40,50]	[30000,40000]	pneumonia
[40,50]	[30000,40000]	gastritis

2-diverse table

+ l-Diversity: Vulnerability

- Suppose that the adversary wants to find out the disease of Bob
- The adversary knows that Bob is unlikely to have breast cancer
- So he knows that Bob is likely to have diabetes

Name	Age	ZIP
Andy	20	10000
Bob	30	20000
Cathy	40	30000
Diane	50	40000

adversary's knowledge

Age	ZIP	Disease
[20,30]	[10000,20000]	breast cancer
[20,30]	[10000,20000]	diabetes
[40,50]	[30000,40000]	pneumonia
[40,50]	[30000,40000]	gastritis

2-diverse table

+ l-Diversity: Other Vulnerabilities

- l-diversity does not consider overall data distribution (*Skewness Attack*)
  - Assume the sensitive attribute is HIV+ or HIV-, and HIV+ is about 1% of the population
  - If one class has 25 HIV+ and 25 HIV-, anyone in the class would be considered to have 50% possibility of being positive, as compared with the 1% of the overall population.
- l-diversity does not consider semantics of sensitive values (*Similarity Attack*)

Zipcode	Age	Salary	Disease
476**	2*	20K	Gastric Ulcer
476**	2*	30K	Gastritis
476**	2*	40K	Stomach Cancer
4790*	≥40	50K	Gastritis
4790*	≥40	100K	Flu
4790*	≥40	70K	Bronchitis
476**	3*	60K	Bronchitis
476**	3*	80K	Pneumonia
476**	3*	90K	Stomach Cancer

+ t-Closeness

- An equivalent class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t
- A table is said to have t-closeness if all equivalence classes have t-closeness

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

+ What Does Attacker Know?

This is against the rules! "flu" is not a quasi-identifier

Caucas	787XX	HIV+	Flu
Asian/AfrAm	787XX	HIV-	Flu
Asian/AfrAm	787XX	HIV+	Shingles
Caucas	787XX	HIV-	Acne
Caucas	787XX	HIV-	Shingles
Caucas	787XX	HIV-	Acne

Table Protected by ℓ-Diversity

+ Differential Privacy

ANDY GREENBERG SECURITY 06.13.16 07:02 PM

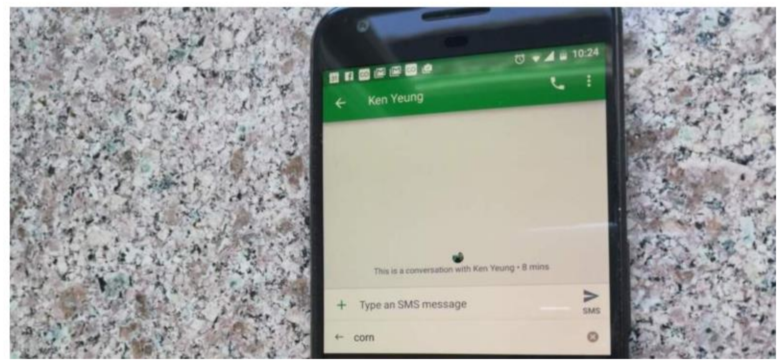
APPLE'S 'DIFFERENTIAL PRIVACY' IS ABOUT COLLECTING YOUR DATA—BUT NOT *YOUR* DATA

Senior vice president of software engineering Craig Federighi.
JUSTIN KANEPS FOR WIRED

APPLE, LIKE PRACTICALLY every mega-corporation, wants to know as much as possible about its customers. But it's also marketed itself as Silicon Valley's privacy champion, one that—unlike so many of its advertising-driven

Following Apple, Google is exploring differential privacy in Gboard for Android

JORDAN NOVET @JORDANNOVET APRIL 6, 2017 6:50 PM



PUBLISHED MAY 18, 2017 IN RESEARCH

NEW TOOLS SAFEGUARD CENSUS DATA ABOUT WHERE YOU LIVE AND WORK

Algorithms guarantee individual privacy without compromising community insights

+ Differential Privacy [Dwork 2006]

- A privacy principle proposed by theoreticians
- More difficult to understand than k-anonymity and l-diversity
- Becomes well-adopted because
  - Its privacy model is without assuming the knowledge the adversary might have.
  - Its definition naturally takes into account algorithm-based attacks

Dwork C, Roth A. The algorithmic foundations of differential privacy[J]. Foundations and Trends in Theoretical Computer Science, 2014, 9(3-4): 211-407.

+ Differential Privacy: Intuition

- Suppose that we have a dataset *D* that contains the medical record of every individual in Australia
- Suppose that Alice is the dataset
- Intuitively, is it OK to publish the following information?
  - Whether Alice has diabetes ✗
  - The total number of diabetes patients in *D* ✓
- Why is it OK to publish the latter but not the former?
- Intuition:
  - The former completely depends on Alice
  - The latter does not depend much on Alice

+ Differential Privacy: Intuition

- In general, we should only publish information that does not highly depend on any particular individual
- This motivates the definition of differential privacy

+ Using Randomized Algorithms

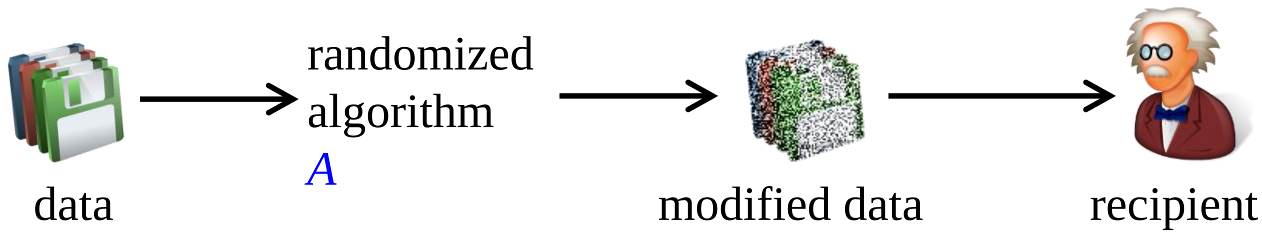
- Differential confidentiality is a process that introduces randomness into the data
- Example: *Are you over 35 years old?*
  - Throw a coin
  - If head, then answer honestly
  - If tail, then throw the coin again and answer "Yes" if head, "No" if tail
- The confidentiality arises from the refutability of the individual responses
- Individual's **deniability is provided via the randomization.**

+ Data Utility

- Data with many responses are significant
  - Positive responses are given to a quarter by people who are under 35 and three-quarters by people who are over 35
  - Given sufficiently large number of responses, can we estimate the true proportion of people over 35 years old (denoted as p), from the observed proportion of people answering “yes” (denoted as q)?

We expect to obtain $q = (1/4)(1-p) + (3/4)p = (1/4) + p/2$ positive responses

+ Differential Privacy: Definition



- Neighboring datasets:
  - Two datasets $\boldsymbol{D}$ and $\boldsymbol{D'}$, such that $\boldsymbol{D'}$ can be obtained by changing one single tuple in $\boldsymbol{D}$

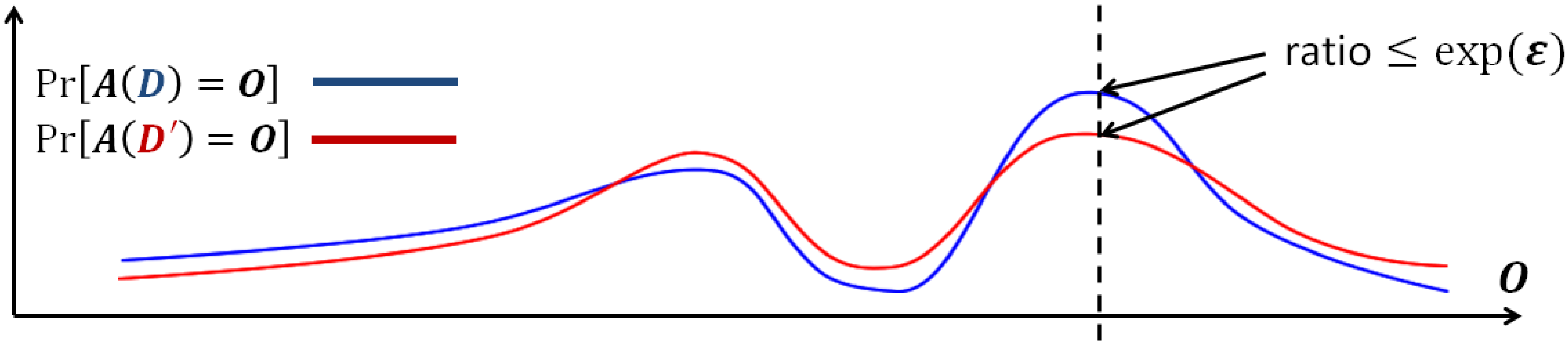
- A randomized algorithm $\boldsymbol{A}$ satisfies $\boldsymbol{\varepsilon}$-differential privacy, iff for any two neighboring datasets $\boldsymbol{D}$ and $\boldsymbol{D'}$ and for any output $\boldsymbol{O}$ of $\boldsymbol{A}$,

$$\Pr[\boldsymbol{A}(\boldsymbol{D}) = \boldsymbol{O}] \leq \exp(\boldsymbol{\varepsilon}) \cdot \Pr[\boldsymbol{A}(\boldsymbol{D'}) = \boldsymbol{O}]$$

- Rationale: The output of the algorithm does not highly depend on any particular tuple in the input

+ Differential Privacy: Illustration

- Illustration of ε-differential privacy



where D and D' are neighboring databases that differ by **at most one** tuple

+ An Example: The Problem

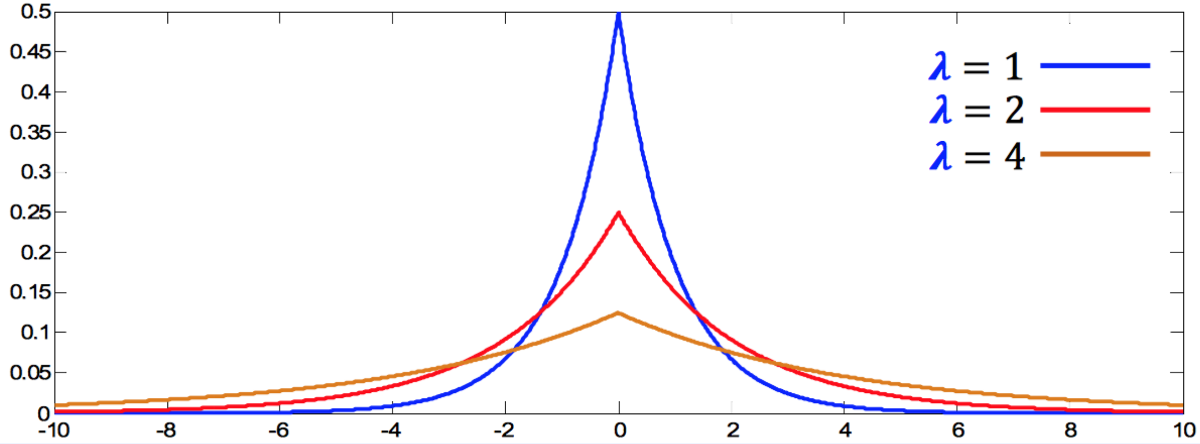
- Suppose that we have a set *D* of medical records
- We want to release statistical information, e.g., the number of diabetes patients in *D:*
- f(D) = select count(*) from *D* where disease = "diabetes";
  - (say we have 1000 diabetes patients in the dataset)

+ Example: How to Release Data

- Non-private solution: Release f(D) directly
- But it violates differential privacy, since
  $$\Pr[\mathrm{f}(\boldsymbol{D}) = \mathbf{1000}] \le \exp(\boldsymbol{\varepsilon}) \cdot \Pr[\mathrm{f}(\boldsymbol{D'}) = \mathbf{1000}]$$
  does not hold
- How to do it in a differentially private manner?
  - Injecting noise into every count before releasing it
    - $\boldsymbol{A}(\boldsymbol{D}, \mathrm{f}) = \mathrm{f}(\boldsymbol{D}) + \mathrm{Noise}$
- Question: what kind of noise should we add?

+ Laplace mechanism

- Noise ~ $Lap(\boldsymbol{\lambda})$, i.e., the noise are i.i.d. random variables drawn from the Laplace distribution with λ



$Lap(\boldsymbol{\lambda})$ λ is referred as the *scale*

+ Sensitivity

- Sensitivity of a function f

$$\Delta f = \max_{D, D'} | f(D) - f(D') |$$

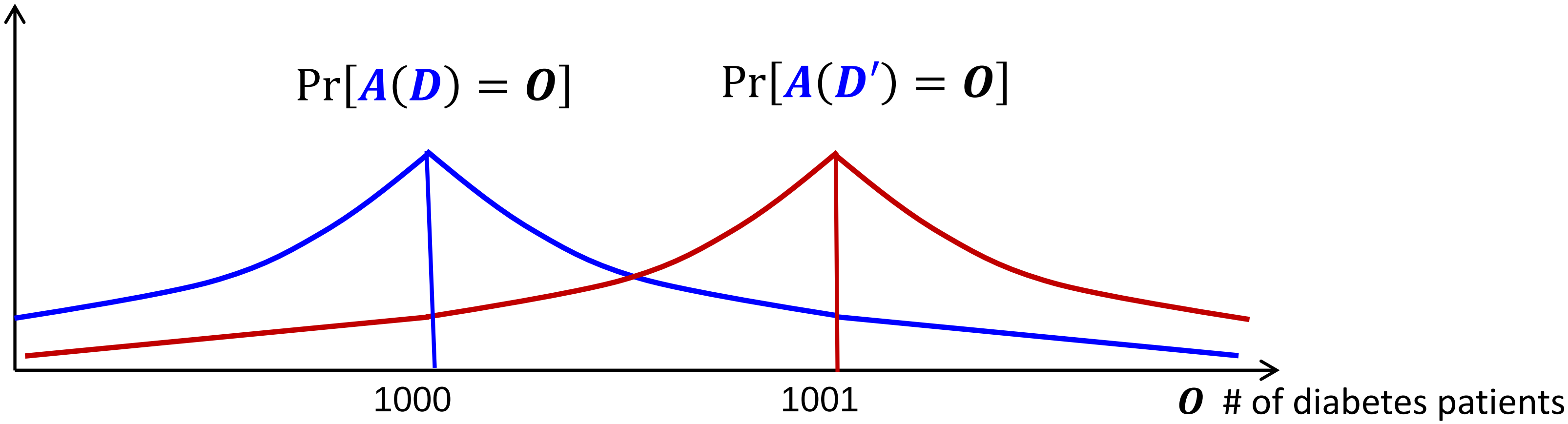
- Sensitivity captures how much one person's data can affect output
- What is sensitivity for counting query? $\Delta f = 1$

+ Adding Laplace Noise

- Add Laplace noise with $\lambda = \Delta f / \varepsilon = 1 / \varepsilon$ before releasing the number of diabetes patients in *D*
  - Noise depends on f and ε, not on the dataset

+ Adding Laplace Noise

- Add Laplace noise with $\lambda = \Delta f / \varepsilon = 1 / \varepsilon$ before releasing the number of diabetes patients in *D*
  - Noise depends on f and ε



+ Statistical Attack Revisited

- Attack queries

```
select count(*)
from   staff
where title = "Professor"
```

```
select sum(salary)
from   staff
where title = "Professor"
```

- Plausible deniability
  - With or without me, you get the same answer, if an ϵ-differential privacy algorithm is used for a sufficiently small ϵ

+ Comparison with k-anonymity, l-diversity and t-closeness

- Differential privacy does not directly model the adversary's knowledge
  - Can achieve ε-DP by adding a random noise value
  - Uncertainty due to noise → plausible deniability
- It is more general
  - There is no restriction on the type of O
  - It can be a number, a table, a set of frequent itemsets, a regression model, etc.

+ Summary

- Data privacy protection methods and limitations
  - k-anonymity
  - l-diversity
  - t-closeness
  - Differential privacy
- We haven't discussed details about the algorithms nor about utilities
  - Note: differential privacy is not an algorithm; it's a definition
- This is a problem that is very important and require much more research

+ Readings

- The EU General Data Protection Regulation (gdpr.eu)
- L Sweeney, "k-anonymity: a model for protecting privacy", *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10 (5), 2002
- A Machanavajjhala, J Gehrke, D Kifer, M Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity", International Conference on Data Engineering (ICDE 2006)
- Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian, "t-Closeness: Privacy beyond k-anonymity and l-diversity" (ICDE 2007)
- C Dwork, "Differential Privacy", International Colloquium on Automata, Languages and Programming (ICALP 2006)