数据隐私方法伦理和实践
Methodology, Ethics and Practice of Data Privacy

2. 隐私定义
Privacy Definition

张兰
中国科学技术大学 计算机学院
2020春季

0. Dimensions of Private Data Usage

Who, What, When, Where, How, Why, How much

Who involved

Data Owners

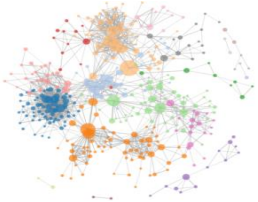
Semi-Honest

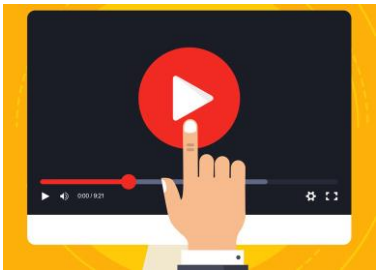
Malicious

- Service Providers
- Other Data Owners
- Attackers
- ...

What to protect

	Name	Age	Gender	Zip Code	Nationality
1	Ann	28	F	13053	Russian
2	Bruce	29	M	13068	Chinese
3	Cary	21	F	13068	Japanese
4	Dick	23	M	13053	American
5	Eshwar	50	M	13053	Indian
6	Fox	55	M	14750	Japanese
7	Gary	47	M	14562	Chinese
8	Helen	49	F	14821	Korean
9	Igor	31	M	13222	American
10	Jean	37	F	13227	American
11	Ken	36	M	13228	American
12	Lewis	35	M	13221	American





What to protect

- **Microdata**: represents a set of records containing information on an individual unit such as a person, a firm, an institution
- **Macrodata**: represents computed/derived statistics
- **Models and patterns**: from machine learning and data mining

What to protect

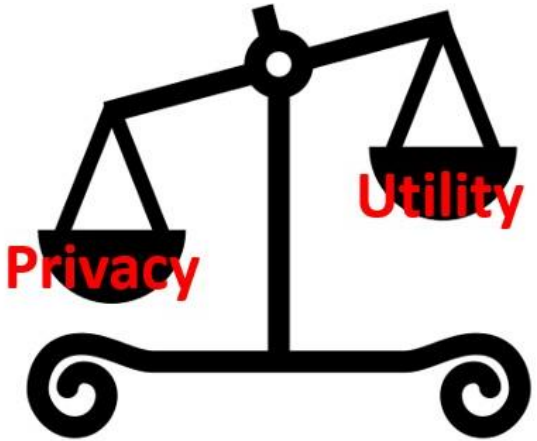
- **Uninformative principle** (Dalenius1977)
  - Access to the published data does not reveal **anything extra** about any target victim, even with the presence of attacker's **background knowledge** obtained from other sources
- **Similar to semantic security of encryption**
  - Knowledge of the ciphertext (and length) of some unknown message does not reveal any additional information on the message that can be feasibly extracted

What to protect

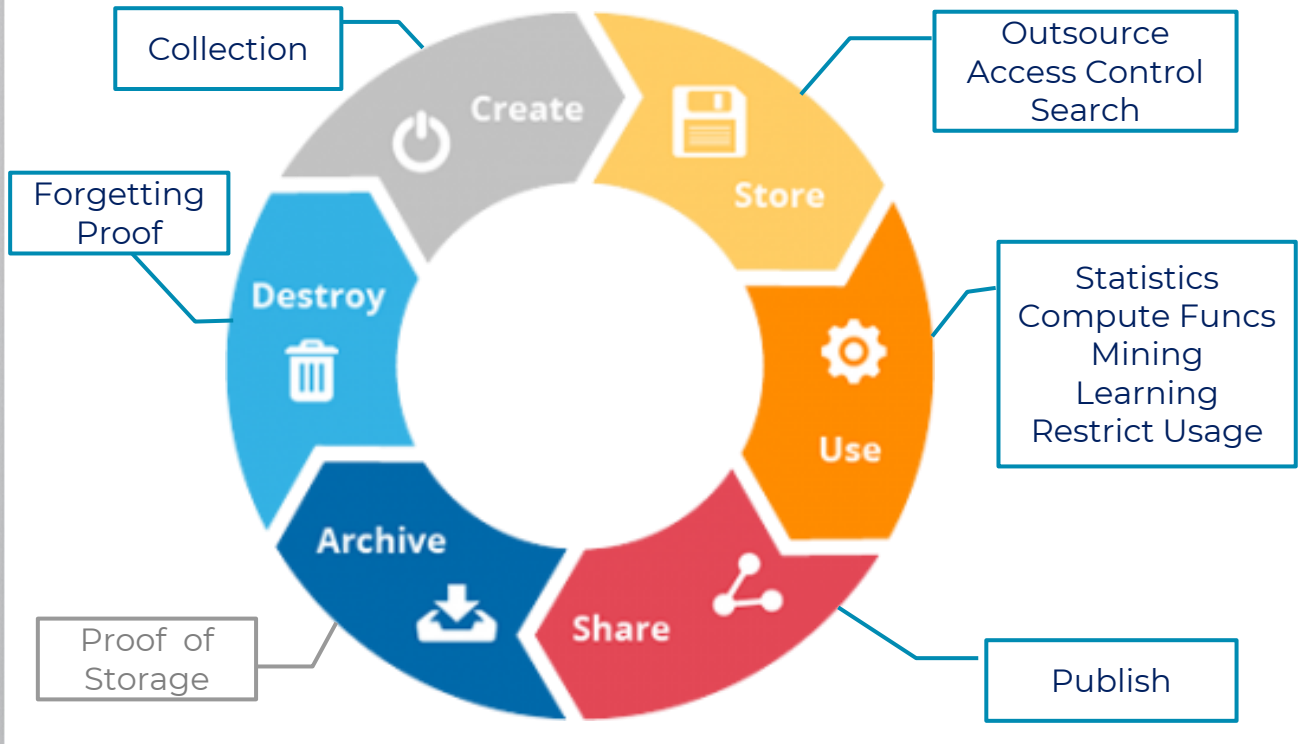
- **Membership disclosure**: Attacker can tell that a given person is in the dataset
- **Identity disclosure**: Attacker can tell which record corresponds to a given person
- **Sensitive attribute disclosure**: Attacker can tell that a given person or record has a certain sensitive attribute

Why so difficult

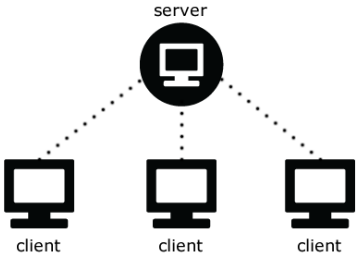
- **Privacy (disclosure risk)**-the risk that a given form of disclosure will arise if the data is released
- **Utility (information loss)**-the information which exist in the initial data but not in released data due to disclosure control methods

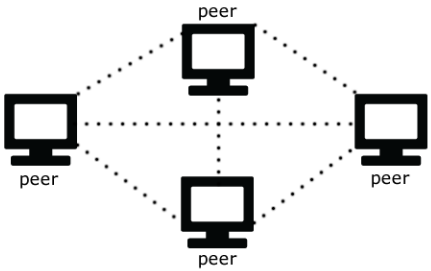


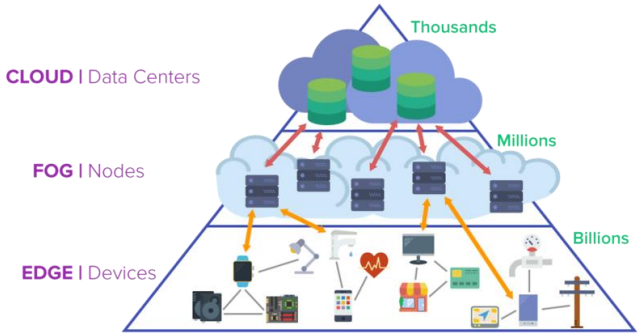
When to protect



Where to operate



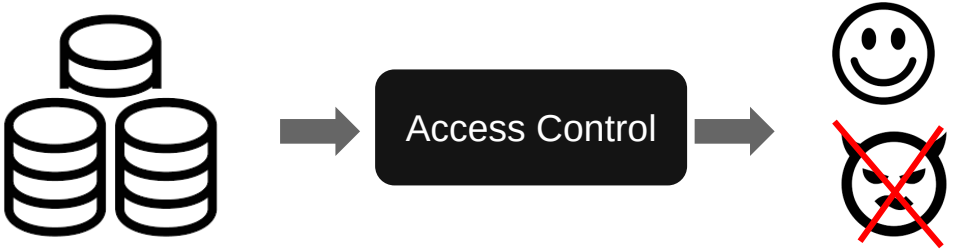




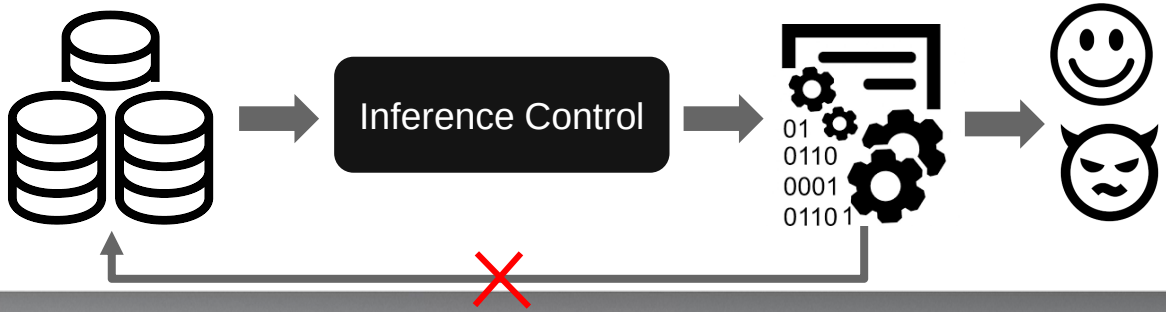
https://www.imagimob.com/blog/edge-computing-needs-edge-ai

How to protect

- **Access control**: protecting information from being accessed by unauthorized users.



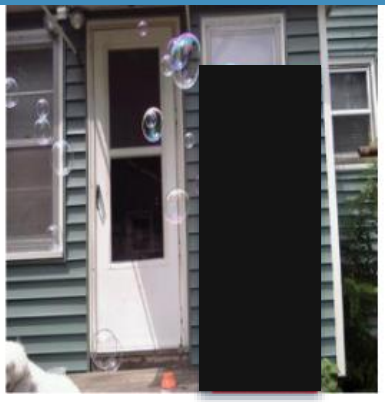
- **Inference control (disclosure control)**: protecting private data from being inferred from sanitized data or models by users

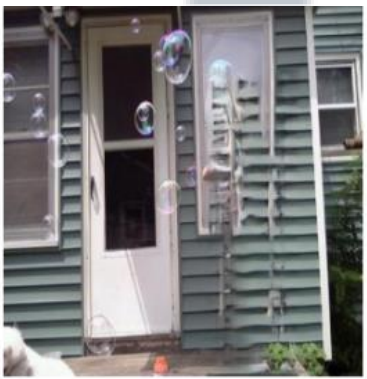


Inference Control Methods

- **不给/部分不给**
  - **Suppress**
  - **Break linkage**
  - **Monitor the queries**: query restriction/auditing
- **给加密的**
  - **Secure Multiparty Computing**
- **给不精准的**
  - **Change granularity**: generalize/aggregate
  - **Change accuracy**: swapping/add noise (DP)
  - **Extract features**: federated learning
- **给合成的(Synthetic data)**

Suppress





Suppress

De-identification

		Age	Gender	Zip Code	Nationality	Condition
1		28	F	13053	Korean	Heart disease
2		29	M	13068	Chinese	Heart disease
3		21	F	13068	Japanese	Viral infection
4		23	M	13053	American	Viral infection
5		50	M	13053	Indian	Cancer
6		55	M	14750	Japanese	Flu
7		47	M	14562	Chinese	Heart disease
8		49	F	14821	Korean	Flu
9		31	M	13222	American	Cancer
10		37	F	13227	American	Cancer
11		36	M	13228	American	Cancer
12		35	M	13221	American	Cancer

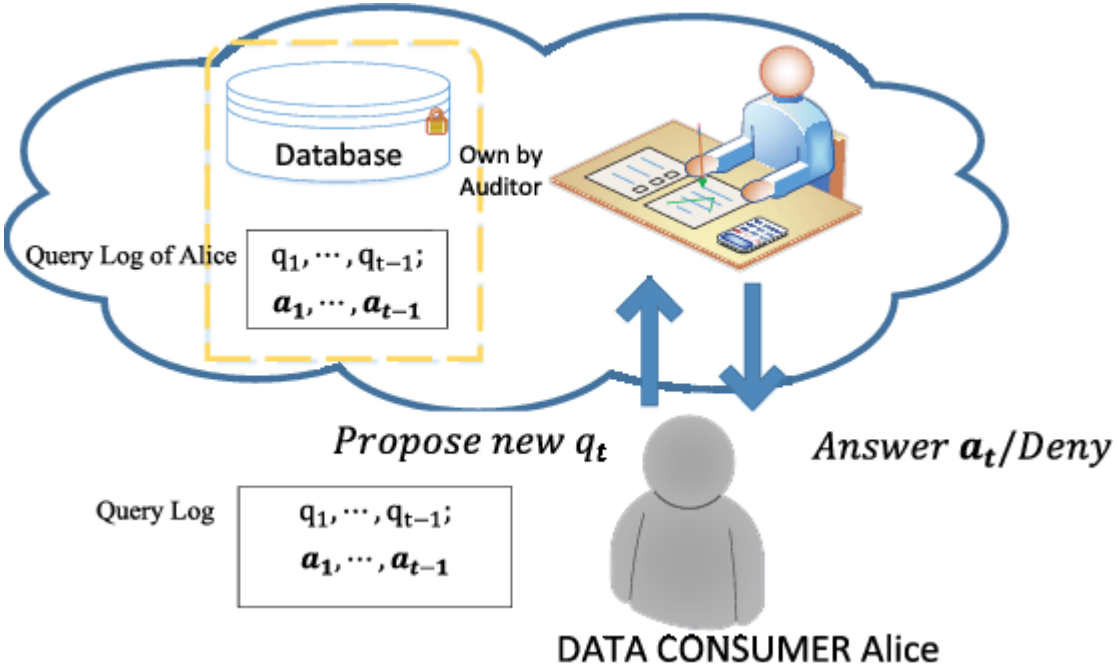
Break Linkage

- **Bucketization**

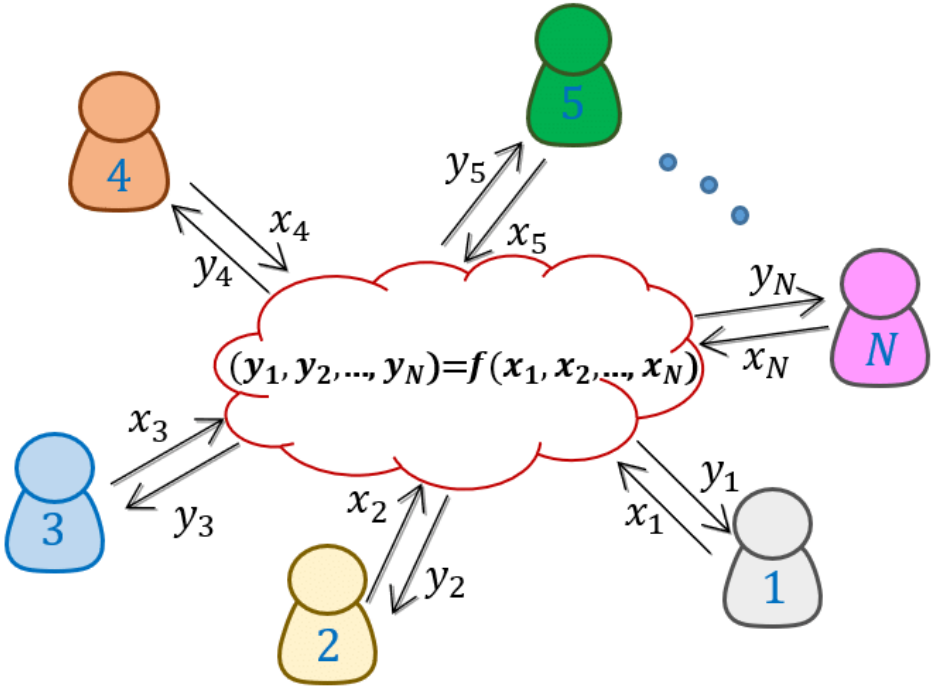
Age	Gender	Zip Code	Nationality	BID
28	F	13053	Korean	**1**
29	M	13068	Chinese	**1**
47	M	14562	Chinese	**1**
•••				
31	M	13222	American	**3**
37	F	13227	American	**3**
36	M	13228	American	**3**
35	M	13221	American	**3**

BID	Condition
1	Heart disease
1	Heart disease
1	Viral infection
•••	
3	Cancer
3	Cancer
3	Cancer
3	Cancer

Monitor the queries: query auditing



Secure Multi-party Computation



Change granularity: generalize/aggregate

	Name	Age	Gender	Zip Code	Nationality	Condition
1	Ann	20-29	Any	130**	Asian	Heart disease
2	Bruce	20-29	Any	130**	Asian	Heart disease
3	Cary	20-29	Any	130**	Asian	Viral infection
4	Dick	20-29	Any	130**	Asian	Viral infection
5	Eshwar	40-59	Any	130**	Asian	Cancer
6	Fox	40-59	Any	14***	Asian	Flu
7	Gary	40-59	Any	14***	Asian	Heart disease
8	Helen	40-59	Any	14***	Asian	Flu
9	Igor	30-39	Any	1322*	American	Cancer
10	Jean	30-39	Any	1322*	American	Cancer
11	Ken	30-39	Any	1322*	American	Cancer
12	Lewis	30-39	Any	1322*	American	Cancer

Change accuracy

» Swapping

	Name	Age	Gender	Zip Code	Nationality	Condition
1	Ann	28	F	13053	Korean	Heart disease
2	Bruce	29	M	13068	Chinese	Heart disease
3	Cary	21	F	13068	Japanese	Viral infection
4	Dick	23	M	13053	American	Viral infection
5	Eshwar	50	M	13053	Indian	Cancer
6	Fox	55	M	14750	Japanese	Flu
7	Gary	47	M	14562	Chinese	Heart disease
8	Helen	49	F	14821	Korean	Flu
9	Igor	31	M	13222	American	Cancer
10	Jean	37	F	13227	American	Cancer
11	Ken	36	M	13228	American	Cancer
12	Lewis	35	M	13221	American	Cancer

Change accuracy

» Add random noise

	Name	Age	Gender	Zip Code	Nationality	Condition
1	Ann	28+2	F	13053	Korean	Heart disease
2	Bruce	29-1	M	13068	Chinese	Heart disease
3	Cary	21+1	F	13068	Japanese	Viral infection
4	Dick	23-2	M	13053	American	Viral infection
•••						

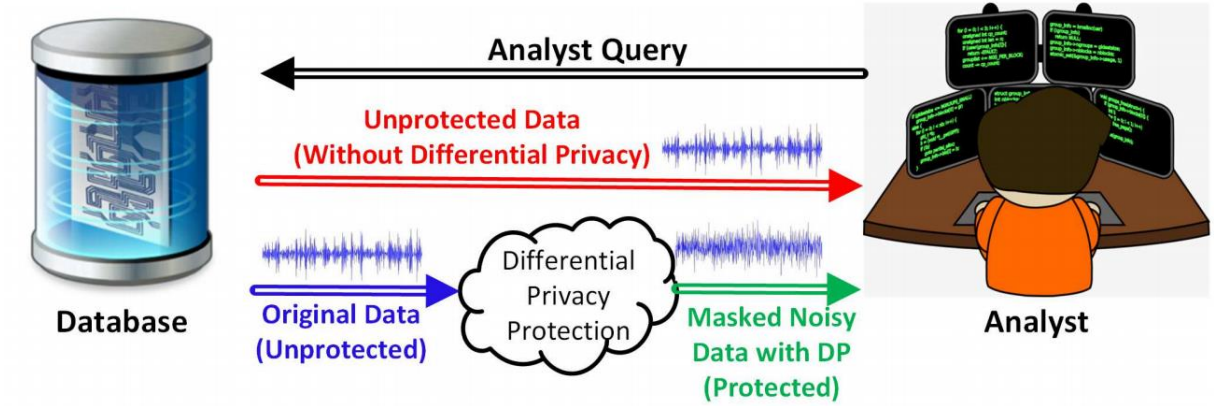
Change accuracy

- **Add (random) noise**

https://www.theguardian.com/world/2020/feb/01/privacy-campaigners-dazzle-camouflage-met-police-surveillance

Change accuracy

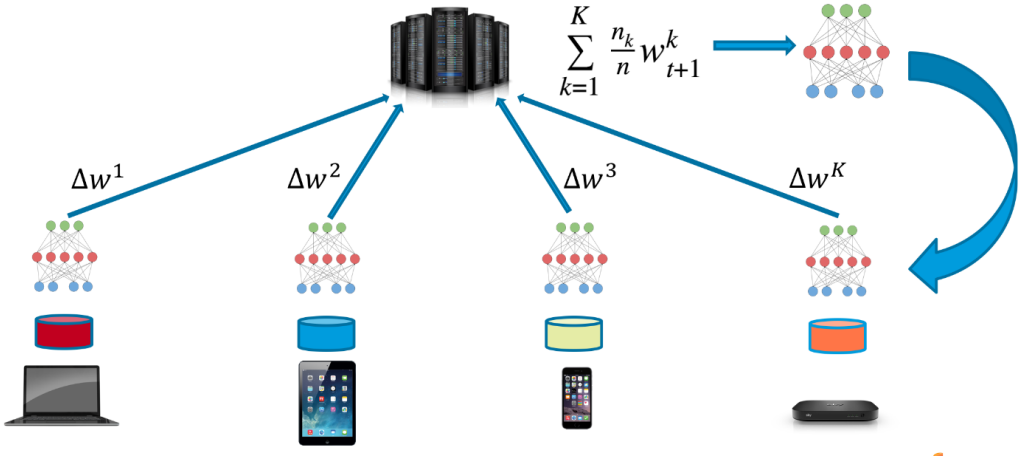
» **Differential privacy**



https://arxiv.org/pdf/1812.02282.pdf

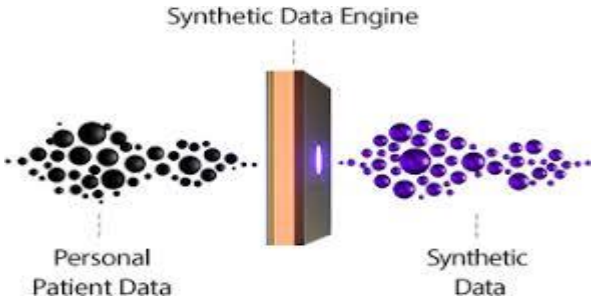
Extract features

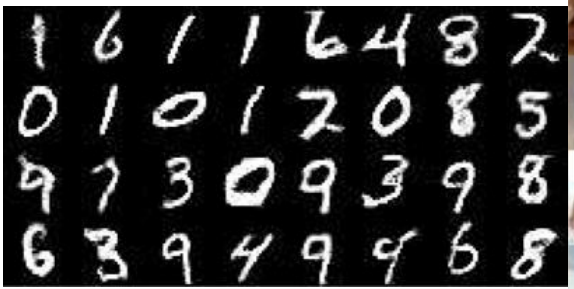
- Federated learning

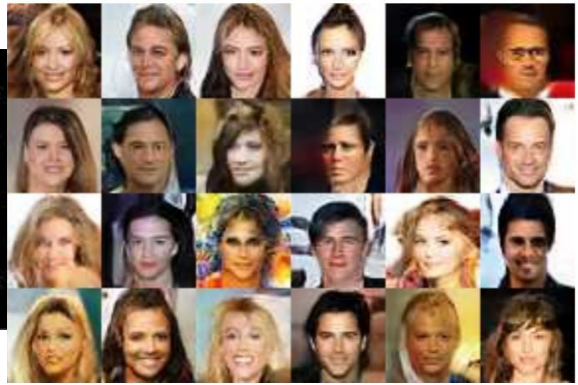


https://miro.medium.com/max/3264/1*HaH611vAy2eB1e42vz3X4g.png

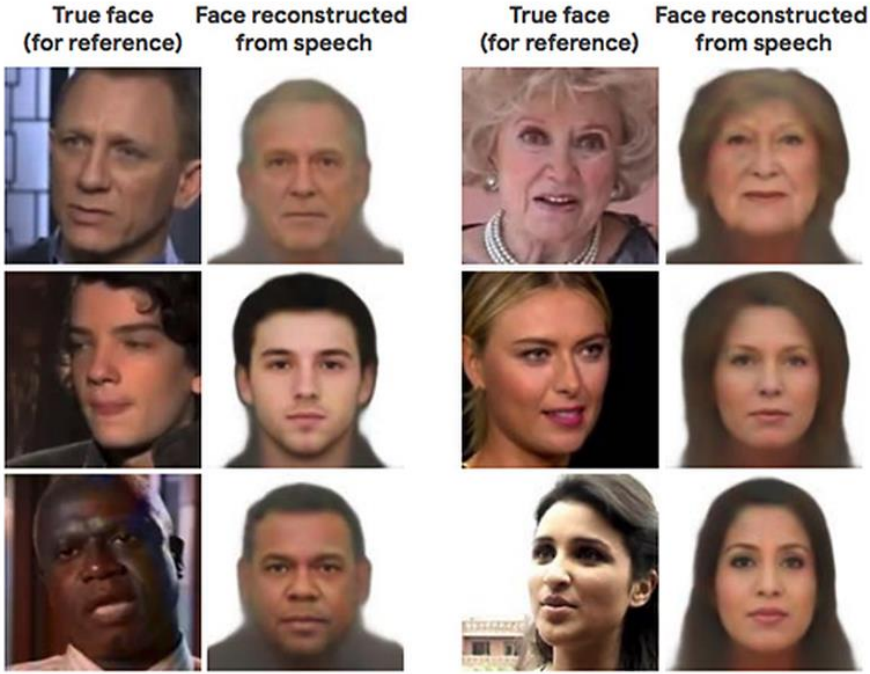
Synthetic data







Synthetic data



https://thenewstack.io/speech2face-reconstructs-faces-using-only-voice-audio/

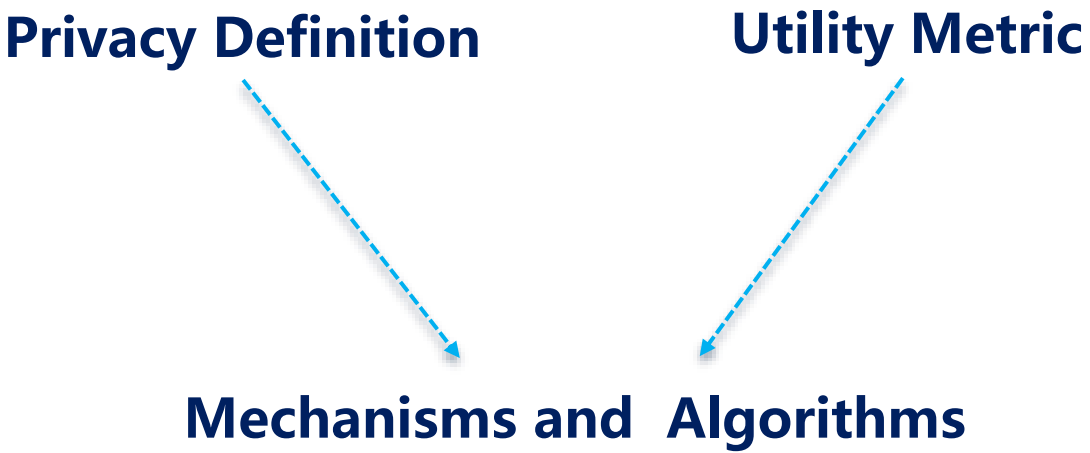
Which one is better?

	Name	Age	Gender	Zip Code	Nationality	Condition
1	Ann	20-29	Any	130**	Asian	Heart disease
2	Bruce	20-29	Any	130**	Asian	Heart disease
3	Cary	20-29	Any	130**	Asian	Viral infection
4	Dick	20-29	Any	130**	Asian	Viral infection
5	Eshwar	40-59	Any	130**	Asian	Cancer

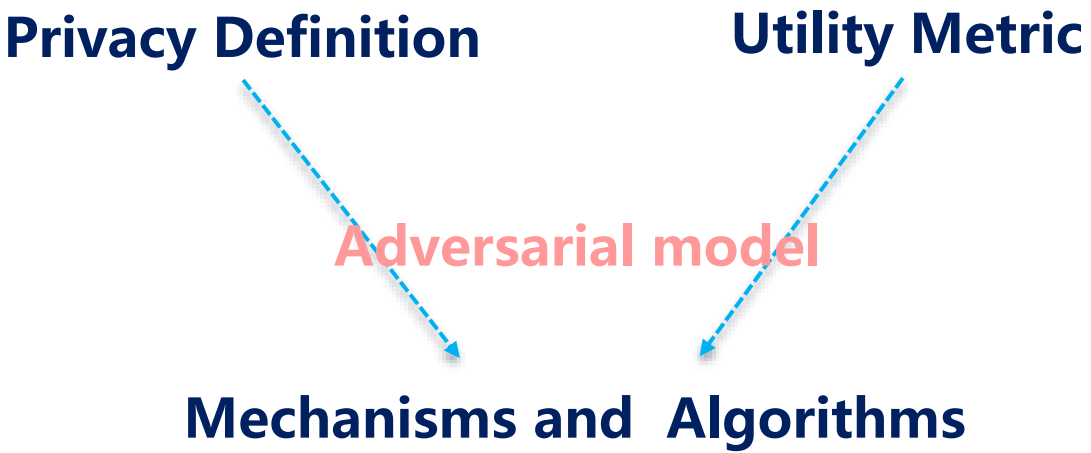
	Name	Age	Gender	Zip Code	Nationality	Condition
1	Ann	28	F	13053	Korean	Heart disease
2	Bruce	29	M	13068	Chinese	Heart disease
3	Cary	21	F	13068	Japanese	Viral infection
4	Dick	23	M	13053	American	Viral infection

	Name	Age	Gender	Zip Code	Nationality	Condition
1	Ann	28+2	F	13053	Korean	Heart disease
2	Bruce	29-1	M	13068	Chinese	Heart disease
3	Cary	21+1	F	13068	Japanese	Viral infection
4	Dick	23-2	M	13053	American	Viral infection
...						

How Much


Privacy Definition
Utility Metric
Mechanisms and Algorithms

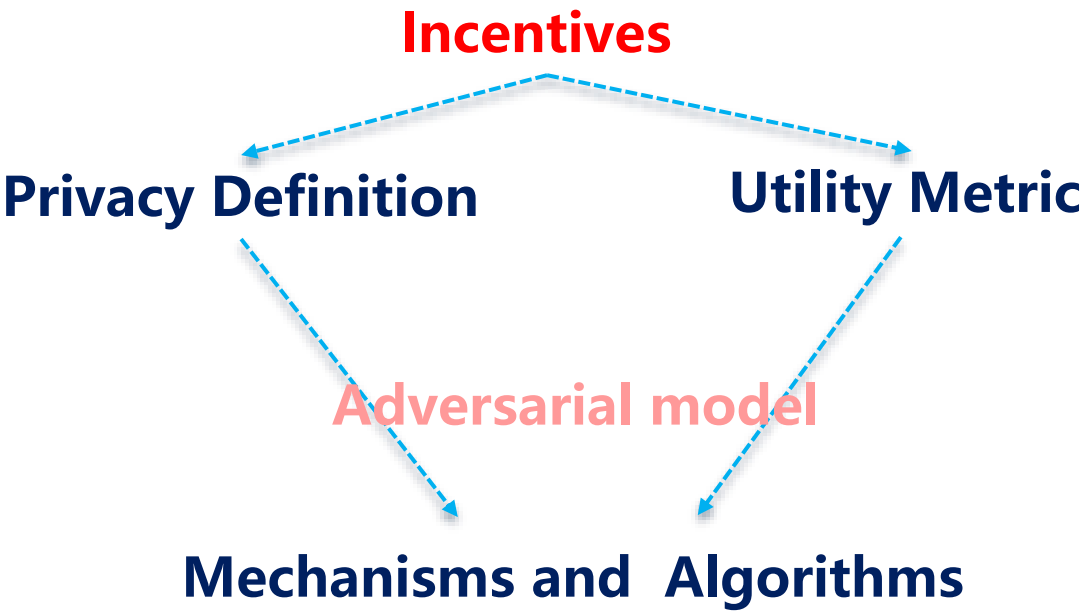
How Much


Privacy Definition
Utility Metric
Adversarial model
Mechanisms and
Algorithms

Adversarial model

- Adversaries:
  - Inside attackers: authorized users with malicious intent
  - Outside attackers: hackers, snoopers
- Adversarial models:
  - Trusted: trusted & tamper-proof
  - Semi-honest: honest but curious
  - Untrusted/malicious: curious + actively modifies data and/or queries
- Background Knowledge
  - The adversary may also have **instance level background knowledge**.
  - The adversary may also know **demographic background** data such as the probability of a condition given an age.

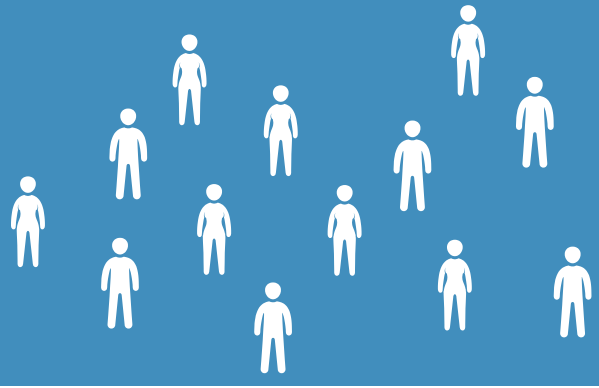
How Much


Incentives
Privacy Definition
Utility Metric
Adversarial model
Mechanisms and Algorithms

“

All the evolution we know of proceeds from the vague to the definite.

— Charles Sanders Peirce



1. K-Anonymity

To protect data from linking attacks,
Samarati and Sweeney proposed k-anonymity.

[1] P. Samarati, “Protecting respondents’ identities in microdata release,” in *Transactions on Knowledge and Data Engineering*, pp. 1010–1027, 2001.
[2] L. Sweeney, “*k*-anonymity: A model for protecting privacy,” *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 10, no. 5, pp. 557–570, 2002.
[3] Sweeney L, Achieving K-Anonymity Privacy Protection using Generalization and Suppression, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, Vol. 10, No. 5, 571-588, 2002.
[4] http://www.cs.emory.edu/~lxiong/cs573/index.html

Basic attempt: de-identification

- Remove/replace “personally identifying information” (PII)
  - PII has no technical meaning or common definition

	Name	Age	Gender	Zip Code	Nationality	Condition
1	Ann	28	F	13053	Korean	Heart disease
2	Bruce	29	M	13068	Chinese	Heart disease
3	Cary	21	F	13068	Japanese	Viral infection
4	Dick	23	M	13053	American	Viral infection
5	Eshwar	50	M	13053	Indian	Cancer
6	Fox	55	M	14750	Japanese	Flu
7	Gary	47	M	14562	Chinese	Heart disease
8	Helen	49	F	14821	Korean	Flu
9	Igor	31	M	13222	American	Cancer
10	Jean	37	F	13227	American	Cancer
11	Ken	36	M	13228	American	Cancer
12	Lewis	35	M	13221	American	Cancer

Basic attempt: de-identification

- Remove/replace “personally identifying information” (PII)
  - PII has no technical meaning or common definition
  - **Linking attack : 87% of the population of the United States can be uniquely identified on the basis of their five-digit zip code, gender, and date of birth.**

Massachusetts GIC Incident

- Massachusetts GIC released “anonymized” data on state employees’ hospital visit
- Then Governor William Weld assured public on privacy

Anonymized

Age	Zip Code	Condition
28	13053	AIDS
29	13068	AIDS
21	13068	Viral infection
23	13053	Viral infection
50	13053	Cancer
55	14750	Flu
47	14562	Heart disease
49	14821	Flu

Massachusetts GIC Incident

- **Re-identification**: then graduate student Sweeney linked the data with Voter registration data in Cambridge and identified Governor Weld's record

Age	Zip Code	Condition
28	13053	AIDS
29	13068	AIDS
21	13068	Viral infection
23	13053	Viral infection
50	13053	Cancer
55	14750	Flu
47	14562	Heart disease
49	14821	Flu

	Age	Zip Code
Ann	28	13053
Cary	21	13068
Dick	23	13053
Helen	49	14821

AOL Query Log Release

- 20 million Web search queries by AOL

AnonID	Query	QueryTime	ItemRank	ClickURL
217	lottery	2006-03-01 11:58:51	1	http://www.calottery.com
217	lottery	2006-03-27 14:10:38	1	http://www.calottery.com
1268	gall stones	2006-05-11 02:12:51		
1268	gallstones	2006-05-11 02:13:02	1	http://www.niddk.nih.gov
1268	ozark horse blankets	2006-03-01 17:39:28	8	http://www.blanketsnmore.com

AOL Query Log Release

- User No. 4417749
  - “numb fingers”, “60 single men”, “dog that urinates on everything”, “landscapers in Lilburn, Ga”, Several people names with last name Arnold, “homes sold in shadow lake subdivision gwinnett county georgia”
  - Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga., frequently researches her friends’ medical ailments and loves her dogs.

The Genome Hacker (2013)

- Yaniv Erlich shows how research participants can be identified from 'anonymous' DNA by cross-referencing their data with publicly available information.
- https://www.nature.com/news/privacy-protections-the-genome-hacker-1.12940

Linkage


BVLGARI


heungjacky
Bvlgari hotel Beijing
58,770 likes
heungjacky Nice stay again at Beijing Bvlgari hotel @bulgarihotels #bvlgari
View all 300 comments
1月28日 · SEE TRANSLATION

Hide in a crowd

https://www.sohu.com/a/251511668_100097557

Hide in a crowd

https://www.sohu.com/a/251511668_100097557

An Example Data Table

	Name	Age	Gender	Zip Code	Nationality	Condition
1	Ann	28	F	13053	Korean	Heart disease
2	Bruce	29	M	13068	Chinese	Heart disease
3	Cary	21	F	13068	Japanese	Viral infection
4	Dick	23	M	13053	American	Viral infection
5	Eshwar	50	M	13053	Indian	Cancer
6	Fox	55	M	14750	Japanese	Flu
7	Gary	47	M	14562	Chinese	Heart disease
8	Helen	49	F	14821	Korean	Flu
9	Igor	31	M	13222	American	Cancer
10	Jean	37	F	13227	American	Cancer
11	Ken	36	M	13228	American	Cancer
12	Lewis	35	M	13221	American	Cancer

Attribute Classification

» Identifier attributes

- Information that leads to a specific entity

	Name	Age	Gender	Zip Code	Nationality	Condition
1	Ann	28	F	13053	Korean	Heart disease
2	Bruce	29	M	13068	Chinese	Heart disease
3	Cary	21	F	13068	Japanese	Viral infection
4	Dick	23	M	13053	American	Viral infection
5	Eshwar	50	M	13053	Indian	Cancer
6	Fox	55	M	14750	Japanese	Flu
7	Gary	47	M	14562	Chinese	Heart disease
8	Helen	49	F	14821	Korean	Flu
9	Igor	31	M	13222	American	Cancer
10	Jean	37	F	13227	American	Cancer
11	Ken	36	M	13228	American	Cancer
12	Lewis	35	M	13221	American	Cancer

Attribute Classification

» Quasi-identifiers (Key attributes)

	Name	Age	Gender	Zip Code	Nationality	Condition
1	Ann	28	F	13053	Korean	Heart disease
2	Bruce	29	M	13068	Chinese	Heart disease
3	Cary	21	F	13068	Japanese	Viral infection
4	Dick	23	M	13053	American	Viral infection
5	Eshwar	50	M	13053	Indian	Cancer
6	Fox	55	M	14750	Japanese	Flu
7	Gary	47	M	14562	Chinese	Heart disease
8	Helen	49	F	14821	Korean	Flu
9	Igor	31	M	13222	American	Cancer
10	Jean	37	F	13227	American	Cancer
11	Ken	36	M	13228	American	Cancer
12	Lewis	35	M	13221	American	Cancer

Attribute Classification

» Quasi-identifiers

- Pieces of information that are not of themselves unique identifiers, but are sufficiently well correlated with an entity that they can be combined with other quasi-identifiers to **create a unique identifier**.
- Age, gender, zip code, and nationality are called quasi-identifier attributes, because by looking at these attributes an adversary may potentially identify an individual in the data set.
- **May be known by an intruder from other sources.**

Attribute Classification

» Confidential or sensitive attributes

- **Assumed to be unknown to an intruder**

	Name	Age	Gender	Zip Code	Nationality	Condition
1	Ann	28	F	13053	Korean	Heart disease
2	Bruce	29	M	13068	Chinese	Heart disease
3	Cary	21	F	13068	Japanese	Viral infection
4	Dick	23	M	13053	American	Viral infection
5	Eshwar	50	M	13053	Indian	Cancer
6	Fox	55	M	14750	Japanese	Flu
7	Gary	47	M	14562	Chinese	Heart disease
8	Helen	49	F	14821	Korean	Flu
9	Igor	31	M	13222	American	Cancer
10	Jean	37	F	13227	American	Cancer
11	Ken	36	M	13228	American	Cancer
12	Lewis	35	M	13221	American	Cancer

Attribute Classification

- **Identifier attributes**
  - Ex: Name and SSN
  - Information that leads to a specific entity
- **Quasi-identifier attributes**
  - Ex: Zip Code and Age
  - May be known by an intruder
- **Confidential or sensitive attributes**
  - Ex: Principal Diagnosis and Annual Income
  - Assumed to be unknown to an intruder

K-Anonymity Definition

- **QI-cluster** – all the tuples with identical combination of quasi-identifier attribute values in that microdata.
- **K-anonymity property** for a masked microdata (MM) is satisfied if every QI-cluster in MM contains k or more tuples.

Definition — K-Anonymity Definition. Given a set of QI attributes, $Q_1, \cdots, Q_d$, release candidate D is said to be k-anonymous with respect to $Q_1, \cdots, Q_d$ if each unique tuple in the projection of D on $Q_1, \cdots, Q_d$ occurs at least k times.

QI-cluster

	Name	Age	Gender	Zip Code	Nationality	Condition
1	Ann	20-29	Any	130**	Asian	Heart disease
2	Bruce	20-29	Any	130**	Asian	Heart disease
3	Cary	20-29	Any	130**	Asian	Viral infection
4	Dick	20-29	Any	130**	Asian	Viral infection
5	Eshwar	40-59	Any	130**	Asian	Cancer
6	Fox	40-59	Any	14***	Asian	Flu
7	Gary	40-59	Any	14***	Asian	Heart disease
8	Helen	40-59	Any	14***	Asian	Flu
9	Igor	30-39	Any	1322*	American	Cancer
10	Jean	30-39	Any	1322*	American	Cancer
11	Ken	30-39	Any	1322*	American	Cancer
12	Lewis	30-39	Any	1322*	American	Cancer

K-Anonymity Definition

Every QI-cluster in MM contains k or more tuples.

	Name	Age	Gender	Zip Code	Nationality	Condition
1	Ann	20-29	Any	130**	Asian	Heart disease
2	Bruce	20-29	Any	130**	Asian	Heart disease
3	Cary	20-29	Any	130**	Asian	Viral infection
4	Dick	20-29	Any	130**	Asian	Viral infection
5	Eshwar	40-59	Any	130**	Asian	Cancer
6	Fox	40-59	Any	14***	Asian	Flu
7	Gary	40-59	Any	14***	Asian	Heart disease
8	Helen	40-59	Any	14***	Asian	Flu
9	Igor	30-39	Any	1322*	American	Cancer
10	Jean	30-39	Any	1322*	American	Cancer
11	Ken	30-39	Any	1322*	American	Cancer
12	Lewis	30-39	Any	1322*	American	Cancer

Example

ID	Age	Zip	Sex	Illness
1	50	41076	Female	AIDS
2	30	41099	Male	Diabetes
3	30	41099	Male	AIDS
4	20	41076	Male	Asthma
5	20	41076	Male	Asthma
6	50	41076	Female	Diabetes
7	50	41076	Female	Tuberculosis

» **K=?**

Example

ID	Age	Zip	Sex	Illness
1	50	41076	Female	AIDS
2	30	41099	Male	Diabetes
3	30	41099	Male	AIDS
4	20	41076	Male	Asthma
5	20	41076	Male	Asthma
6	50	41076	Female	Diabetes
7	50	41076	Female	Tuberculosis

» **K=2**

Achieving k-Anonymity: k-anonymization

- **Generalization**
  - Replace specific quasi-identifiers with more general values until get k identical values

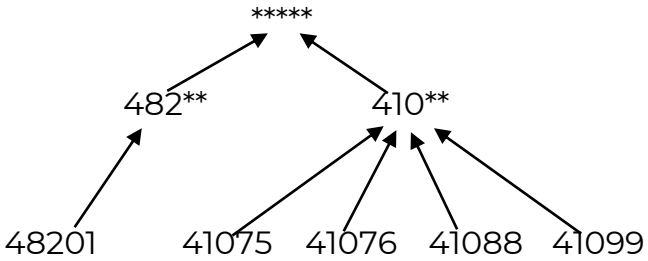
    Example: area code instead of phone number
  - Partition ordered-value domains into intervals
- **Suppression**
  - When generalization causes too much information loss
- **Lots of algorithms in the literature**
  - Aim to minimize various form of information loss

Domain and Value Generalization Hierarchies

Z_2 = {*****}

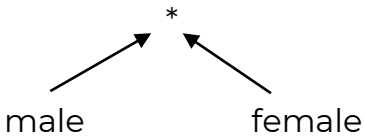
Z_1 = {482**, 410**}

Z_0 = {48201, 41075, 41076, ...}



S_1 = {*}

S_0 = {male, female}



[Samarati 2001, Sweeney 2002]

Generalization Types

- All Attributes:
  - **Full domain generalization** [Samarati 2001, LeFevre 2006] Map the entire domain of each quasi-identifier attribute to a more general domain in its domain generalization hierarchy. Guarantee that all values of a particular attribute in V belong to the same domain.
  - **Iyengar generalization** [Iyengar 2002] A more flexible scheme, which also uses a fixed VGH, but allows different values of an attribute to be generalized to different levels.
  - **Cell-level generalization** [Lunacek 2006]
- Numerical Attributes
  - **Predefined hierarchy** [Iyengar 2002]
  - **Computed hierarchy** [LeFevre 2006]

2-Anonymity

Tuple	Age	ZipCode	Sex
r_1	50	41076	Male
r_2	30	41075	Female
r_3	30	41099	Female
r_4	20	48201	Male
r_5	20	41075	Male

Full domain generalization

Tuple	Age	ZipCode	Sex
r_1	20-30	*****	Male
r_2	20-30	*****	Male
r_3	30-40	*****	Female
r_4	30-40	*****	Female
r_5	30-40	*****	Female

Cell-level generalization

Tuple	Age	ZipCode	Sex
r_1	20-30	410**	Male
r_2	20-30	410**	Male
r_3	30-40	*****	Female
r_4	30-40	*****	Female
r_5	30-40	*****	Female

Attacks Against K-Anonymity

- **Unsorted Matching Attack**
  - This attack is based on the order in which tuples appear in the released table. [Sweeney 2002]
  - Solution: Randomly sort the tuples before releasing.
- **Temporal Attack**
  - Adding or removing tuples may compromise k-anonymity protection [Sweeney 2002].

Attacks Against K-Anonymity

» Complementary Release Attack

- Different releases can be linked together to compromise k-anonymity [Sweeney 2002].
- Solution: Consider all of the released tables before release the new one, and try to avoid linking.
- Other data holders may release some data that can be used in this kind of attack. Generally, this kind of attack is hard to be prohibited completely.

Multi-view release

» To increase data utility, the data publisher may release multiple views of a single original data set, where the released views are outputs of one (or more) of the above sanitization mechanisms.

(a) Marginal on gender, nationality

Gender	Nationality	Count
F	Russian	1
F	Japanese	1
F	Korean	1
F	American	1
M	Chinese	2
M	American	4
M	Indian	1
M	Japanese	1

(b) Marginal on gender, condition

Gender	Condition	Count
F	Heart disease	1
F	Viral infection	1
F	Flu	1
F	Cancer	1
M	Heart disease	2
M	Viral infection	1
M	Flu	1
M	Cancer	4

Attacks Against K-Anonymity

- A 3-anonymous patient table

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

Attacks Against K-Anonymity

Homogeneity Attack

Bob	
Zipcode	***Age***
47678	27

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

Attacks Against K-Anonymity

Homogeneity Attack

Bob	
Zipcode	***Age***
47678	27

Background Knowledge Attack

Carl (Japanese)	
Zipcode	***Age***
47673	36

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

Attacks Against K-Anonymity

- K-Anonymity protects against identity disclosure but not provide sufficient protection against attribute disclosure

- **K-Anonymity does not provide privacy if**
  - Sensitive values in an equivalence class lack diversity [Truta 2006, Machanavajjhala 2006].
  - The attacker has background knowledge [Machanavajjhala 2006].

P-Sensitive K-Anonymity Definition

» **P-sensitive K-anonymity property**

A MM satisfies p-sensitive k-anonymity property if it

- satisfies k-anonymity and
- the number of distinct attribute values for each sensitive attribute is at least p within the same QI-cluster from the MM [Truta 2006].

P-Sensitive K-Anonymity

» A 3-anonymous patient table

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

» **P=?**

P-Sensitive K-Anonymity

» A 3-anonymous patient table

Zipcode	Age	Disease
476**	<40	Heart Disease
476**	<40	Heart Disease
476**	<40	Heart Disease
4790*	≥40	Flu
4790*	≥40	Heart Disease
4790*	≥40	Cancer
476**	<40	Heart Disease
476**	<40	Cancer
476**	<40	Cancer

» P=?

Revisiting the example

~~Homogeneity Attack~~

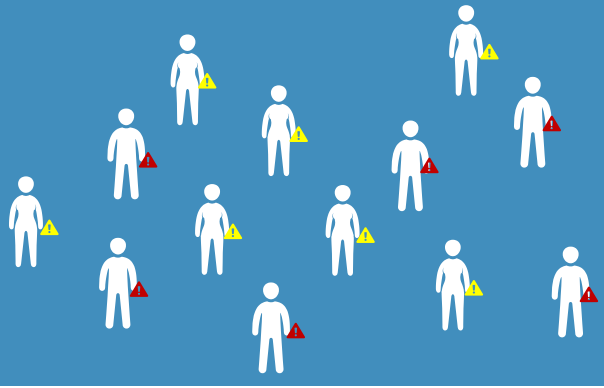
Bob	
Zipcode	***Age***
47678	27

Background Knowledge Attack

Carl (Japanese)	
Zipcode	***Age***
47673	36

Zipcode	Age	Disease
476**	<40	Heart Disease
476**	<40	Heart Disease
476**	<40	Heart Disease
4790*	≥40	Flu
4790*	≥40	Heart Disease
4790*	≥40	Cancer
476**	<40	Heart Disease
476**	<40	Cancer
476**	<40	Cancer

2. l-Diversity



To protect data from attackers when the sensitive information lacks diversity.

[1] Machanavajjhala, Ashwin, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. "l-diversity: Privacy beyond k-anonymity." ACM Transactions on Knowledge Discovery from Data (TKDD) 1, no. 1 (2007): 3-es

l-Diversity

- **Protect against attribute disclosure**
- **Each QI cluster has at least l well-represented sensitive values.**
- **1. Distinct l-diversity**
  - There are at least l-distinct sensitive values.
  - Doesn't prevent the probabilistic inference attacks

…	Cancer
…	Heart Disease
…	Cancer
…	Cancer
…	Cancer
…	Cancer
…	Cancer

90% records have cancer

2. Entropy l-diversity

- Each QI cluster not only must have enough different sensitive values, but also the **different sensitive values must be distributed evenly enough**.

- The entropy of the distribution of sensitive values in each QI cluster is **at least log(l)**.

- Sometimes this maybe too restrictive. When some values are very common, the entropy of the entire table may be very low.

3. Recursive (c,l)-diversity

» **The most frequent value does not appear too frequently, while less common values are ensured to not appear too infrequently.**

- Let $s_1, \ldots, s_m$ be the possible values of the sensitive attribute S in a group of tuples with generalized QI value q*.
- We sort the counts $n(q^*,s_1), \ldots, n(q^*,s_m)$ in descending order and name the elements of the resulting sequence $r_1, \ldots, r_m$.
- (c, 2)-diverse: $r_1 < c(r_2+r_3+\ldots+r_m)$ should be true, where r_i is frequency.
- (c, l)-diversity: $r_1 < c(r_l+r_{l+1}+\ldots+r_m)$ should be true.
- The recursive (*c, l*)-diversity, thus, can be interpreted in terms of adversarial background knowledge. It guards against all adversaries who possess at most ***l− 2*** statements of the form "Bob does not have heart disease".

» Negative/Positive Disclosure-Recursive (c_1, c_2, l)-Diversity

Positive & Negative Disclosure

- **Positive Disclosure**: Publishing the table $T^\star$ that was derived from T results in a positive disclosure if the adversary can **correctly identify** the value of a sensitive attribute with high probability

- **Negative disclosure**: Publishing the table $T^\star$ that was derived from T results in a negative disclosure if the adversary can **correctly eliminate** some possible values of the sensitive attribute with high probability

Limitations of l-Diversity & p-sensitive k-anonymity

- Attribute disclosure not completely prevented.
  - **Skewness Attack [Li 2007]**
- Two sensitive values
  - HIV positive (1%) and HIV negative (99%).
- Serious privacy risk
  - Consider an equivalence class that contains a large number of positive records compared to negative records.
- **l-diversity & p-sensitive k-anonymity does not differentiate**
  - Equivalence class 1: 49 positive + 1 negative.
  - Equivalence class 2: 1 positive + 49 negative.

Neither Necessary, Nor Sufficient

Original Data

…	HIV-
…	HIV-
…	HIV-
…	HIV-
…	HIV+
…	HIV-
…	HIV-
…	HIV-
…	HIV-
…	HIV-

99% have HIV-

Anonymization A

Q1	HIV+
Q1	HIV-
Q1	HIV+
Q1	HIV-
Q1	HIV+
Q1	HIV-
…	HIV-
…	HIV-
…	HIV-
…	HIV-

Diverse, but leak a lot of information

Anonymization A

Q1	HIV-
Q1	HIV-
Q1	HIV-
Q1	HIV-
Q1	HIV+
Q1	HIV-
…	HIV-
…	HIV-
…	HIV-
…	HIV-

Not diverse, but not leak anything

Similarity Attack

Bob	
Zip	***Age***
47678	27

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

Conclusion

- Bob's salary is in [20k,40k], which is **relative low**.
- Bob has some **stomach-related disease.**

Semantic meanings of sensitive values not considered.

Multiple Sensitive Attributes

- Previous discussions only addressed single sensitive attributes.
- Suppose S and V are two sensitive attributes, and consider the QI cluster with the following tuples:

Q	S_1	V_1
Q	S_1	V_2
Q	S_2	V_3
Q	S_3	V_3

- To address this problem we can add the additional sensitive attributes to the quasi-identifier.

Bayes-Optimal privacy

- The attributes in the input table are considered to be partitioned into non-sensitive QI attributes (called Q) and sensitive attributes (called S). **The adversary is assumed to know the complete joint distribution f of Q and S.**

- Publishing a generalized table breaches privacy if the adversary's **prior belief** in an individual's sensitive attribute **is very different from** the adversary's **posterior belief** after seeing the published generalized table.

Bayes-Optimal privacy

- Adversary Alice's prior belief $\alpha(q,s)$ is her background knowledge:

$$\alpha(q,s) = P_f(t[S] = s | t[Q] = q) = \frac{f(s,q)}{\sum_{s' \in S} f(s',q)}$$

- On observing the published table T^* which is generalized from T, and in which Bob's quasi-identifier q has been generalized to $q*$, her posterior belief about Bob's sensitive attribute is $\beta(q,s,T^*)$:

$$\beta(q,s,T^*) = P_f(t[S] = s | t[Q] = q \text{ and } T^* \text{ and } t \in T^*)$$

Bayes-Optimal privacy

- Privacy is measured by the information gain of an observer.

  Information Gain = Posterior Belief – Prior Belief

$$\alpha(q,s) = P_f(t[S] = s | t[Q] = q) = \frac{f(s,q)}{\sum_{s' \in S} f(s',q)} \qquad \beta(q,s,T^*) = \frac{n(q*,s)\frac{f(s|q)}{f(s|q*)}}{\sum_{s' \in S} n(q*,s')\frac{f(s'|q)}{f(s'|q*)}}.$$

- Publishing a table T satisfies Bayes-Optimal privacy if the distance between $\alpha(q,s)$ and $\beta(q,s,T)$ is small for every q $\in$ Q and for every s $\in$ S; where distance is measured either using the difference or ratio of the two quantities.

Bayes-Optimal privacy

- Formal but impractical.
  - The data publisher is unlikely to know the full distribution f.
  - It is unlikely that the adversary knows the entire joint distribution either.
  - The data publisher may not know the exact knowledge the adversary possesses.
  - The above analysis captures only distributional knowledge and does not capture instance level knowledge.
  - There will be multiple adversaries. The data publisher would have to be able to specify which of these adversaries are guarded against.
  - Checking the Bayes-Optimal condition for every (q, s) combination in the domain might be computationally tedious.

3. t-Closeness

To protect data from attackers with domain knowledge.

[1] Li, Ninghui, Tiancheng Li, and Suresh Venkatasubramanian. "t-closeness: Privacy beyond k-anonymity and l-diversity." In 2007 IEEE 23rd International Conference on Data Engineering, pp. 106-115. IEEE, 2007.

Information Gain

476**	20-30	negative
476**	20-30	negative
476**	20-30	negative
476**	20-30	**positive**
476**	20-30	negative
479**	31-40	negative
479**	31-40	**positive**
479**	31-40	negative
479**	31-40	**positive**
479**	31-40	**positive**
...	...	...

- **Information Gain = Posterior Belief (S|Q)– Prior Belief(S|Q&T*)**
- **B_0**: Alice thinks that Bob may has the disease because he is acting sick
- **B_1**: Based on **P_{T*}**, only 1% population has this disease. She believes that Bob is in that one percent.
- **B_2**: Alice knows Bob's age, so he is in cluster 2. **P_s** is the distribution of cluster 2.
- Based on P she decides that it is quite likely that Bob has the disease.

t-Closeness

- l-diversity requires that P has diversity.
- We want to limit the gain between B_0 and B_2.
- If the change between B_0 and B_1 is large, means that P_{T*} contains lots of new information. But P_{T*} is public. So we focus on the gain between B_1 and B_2.The closer P_s and P_{T*} are, the closer B_{01} and B_2 are.
- **t-Closeness**: distribution of sensitive attributes within each quasi-identifier group should be "close" to their distribution in the entire original database.
- A QI cluster is said to have t-closeness
  - **If the distance between P_s and P_{T*} is no more than a threshold t**

Distance of Two Probabilistic Distributions

- Two distributions
  - $P = (p_1, p_2, \dots, p_m), Q = (q_1, q_2, \dots, q_m)$
- Variational distance

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

- Kullback-Leibler(KL) divergence

$$D[P, Q] = \sum_{i-1}^{m} p_i log \frac{p_1}{q_i} = H(P) - H(P, Q)$$

Distance of Two Probabilistic Distributions

- *Q* = {3k,4k,5k,6k,7k,8k,9k,10k,11k}
- *P1* = {3k,4k,5k}
- *P2* = {6k,8k,11k}

- *P1* has more information leakage than *P2* because there are fewer people in that salary range and thus they are easier to identify, thus we should have D[*P1*,Q]>D[*P2*,Q].

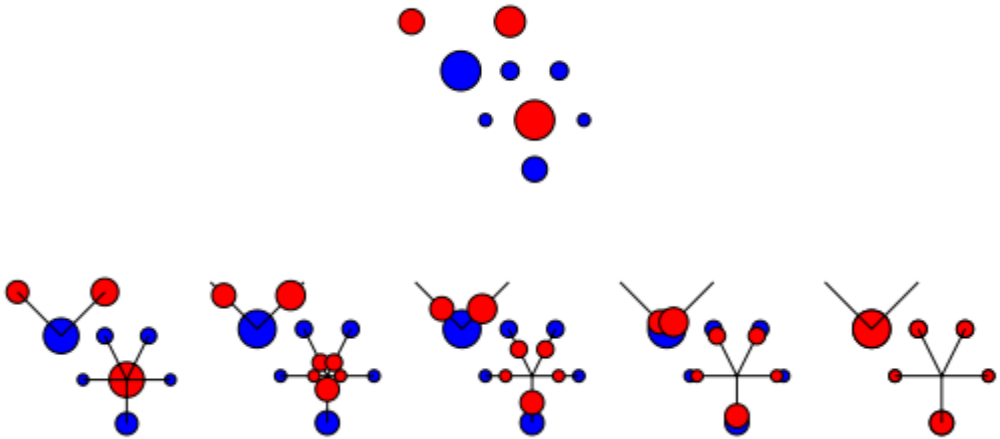
- However, these algorithms just view 3k and 6k as different points and don't attach semantic meaning to them. They would calculate this wrong.

Distance of Two Probabilistic Distributions

- Earth Mover's Distance
  - Suppose that several suppliers, each with a given amount of goods, are required to supply several consumers, each with a given limited capacity. For each supplier-consumer pair, the cost of transporting a single unit of goods is given. The transportation problem is then to find a least-expensive flow of goods from the suppliers to the consumers that satisfies the consumers' demand.
  - Similarly, here the problem is transforming one distribution to another with minimum work done.

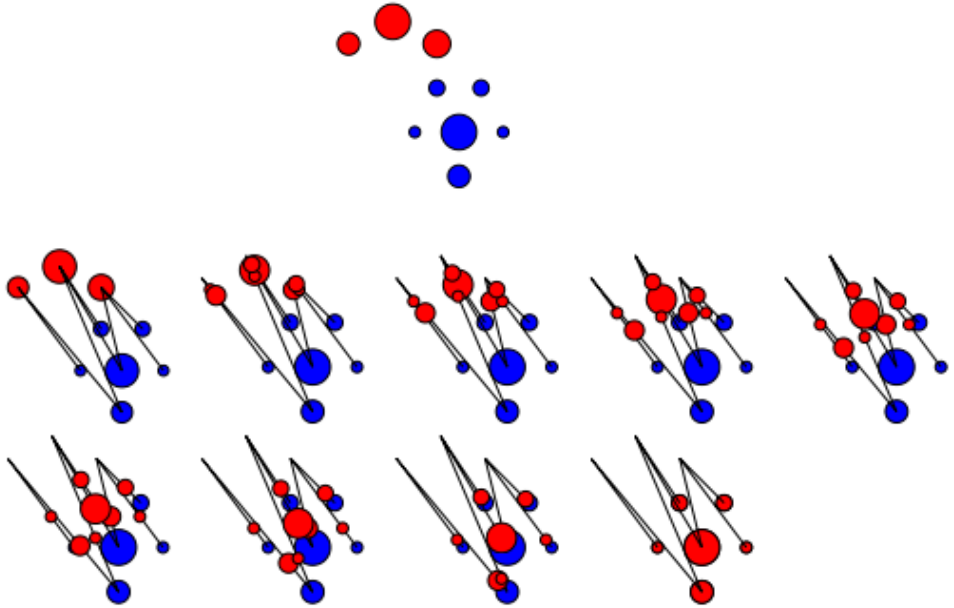
Distance of Two Probabilistic Distributions

- Earth Mover's Distance



Distance of Two Probabilistic Distributions

- Earth Mover's Distance



Distance of Two Probabilistic Distributions

- Earth Mover's Distance

  There constraints guarantee that P is transformed to Q by the mass flow F. Once the transportation problem is solved, the EMD is defined to be the total work.

$$D[P,Q] = WORK(P,Q,F) = \sum_{i=1}^{m}\sum_{j=1}^{m} d_{ij}f_{ij}$$

$$f_{ij} \geq 0, 1 \leq i \leq m, 1 \leq j \leq m \quad (c1)$$

$$p_i - \sum_{j=1}^{m} f_{ij} + \sum_{i=1}^{m} f_{ji} = q_i \quad (c2)$$

$$\sum_{i=1}^{m}\sum_{j=1}^{m} f_{ij} = \sum_{i=1}^{m} p_i = \sum_{i=1}^{m} q_i = 1 \quad (c3)$$

Distance of Two Probabilistic Distributions

- **Numerical Distance**
  - Ordered Distance: Distance between 2 numerical attributes (e.g. age) is based on the number of values between them in the total order.
- **Categorical Distance**
  - Equal distance: for categorical attributes (ie diseases), order does not always matter. We can either view the ground distance between 2 categorical attributes as always being 1.
  - Hierarchical distance: H is the height of the domain hierarchy. The distance between two leaves is defined to be level(v1,v2)/H where level (v1,v2) is the height of the lowest common ancestor node.

Properties of t-closeness

- **Generalization property**: If A and B are generalizations on the table T such that A is more general than B and T satisfies t-closeness using B, then T also satisfies t-closeness using A.

- **Subset property**: If C is a set of attributes in the table T and if T satisfies t-closeness with respect to C, then T also satisfies t-closeness with respect to any set of attributes D such that D is a subset of C.

k-Anonymous, l-diverse, t-close Dataset

» Is it secure?

476**	HIV-	Heart Disease
476**	HIV-	Flu
476**	HIV-	Cancer
476**	HIV-	Flu
476**	HIV+	Heart Disease
479**	HIV-	Cancer
479**	HIV-	Heart Disease
479**	HIV+	Flu
479**	HIV-	Heart Disease
479**	HIV-	Flu

Issues with Syntactic Privacy Notions

- **Syntactic**
  - Focuses on data transformation, not on what can be learned from the anonymized dataset
  - “k-anonymous” dataset (or its variants) can leak sensitive information
- **“Quasi-identifier” fallacy**
  - Assumes a priori that attacker will not know certain information about his target
  - Any attribute can be a potential quasi-identifier (AOL example)

4. Definitions with Background Knowledge

What the adversary knows.

[1] A. Evfimievski, J. Gehrke, and R. Srikant, “Limiting privacy breaches in privacy-preserving data mining,” in *Proceedings of the 22nd ACM SIGMODSIGACT-SIGART Symposium on Principles of Database Systems (PODS)*, 2003.
[2] D. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Halpern, “Worst case background knowledge for privacy preserving data publishing,” in Proceedings of the IEEE International Conference on Data Engineering (ICDE), 2007.
[3] B.-C. Chen, K. LeFevre, and R. Ramakrishnan, “PrivacySkyline: Privacy with multidimensional adversarial knowledge,” in Proceedings of the 33rd International Conference on Very Large Databases (VLDB), 2007.
[4] C. Dwork, “Differential privacy,” in ICALP, 2006.

Questions

- How does a data publisher decide which attributes should be included in the set of QI attributes?
- QI attributes are just a special case of background knowledge.
- k-anonymity considered background knowledge quantified by the set of QI attributes.
- l-Diversity considered adversaries possessing negation statements in addition to QI attributes.
  - "Bob does not have heart disease"
- t-Closeness considered adversaries having some prior belief in an individual's sensitive attribute.
  - "Bob has 30% probability of having heart disease"

Background Knowledge

- In general, one can describe background knowledge using **Boolean logic sentences** and seek to provide privacy protection against an adversary who knows a certain number of such sentences.

- To understand whether $D*$ is safe for release, we consider an adversary whose goal is to predict whether a target individual *t* has a target sensitive value *s*. The adversary has access to the **release candidate *D**** , as well as **his own knowledge *K***.

- A robust privacy criterion should place an **upper bound** on the adversary's confidence in predicting any individual *t* to have sensitive value *s*.

- Breach probability: $\max_{t,s} P(t \ has \ s | K, D*) < c$

Boolean Background Knowledge

- Pr(Eshwar has Cancer |D∗) = 1/4
- K: Eshwar doesn't have flu:
  - Pr(Eshwar has Cancer |K,D∗) = 1/2

	Name	Age	Gender	Zip Code	Nationality	Condition
1	Ann	20-29	Any	130**	Asian	Heart disease
2	Bruce	20-29	Any	130**	Asian	Heart disease
3	Cary	20-29	Any	130**	Asian	Viral infection
4	Dick	20-29	Any	130**	Asian	Viral infection
5	Eshwar	40-59	Any	130**	Asian	Cancer
6	Fox	40-59	Any	14***	Asian	Flu
7	Gary	40-59	Any	14***	Asian	Heart disease
8	Helen	40-59	Any	14***	Asian	Flu
9	Igor	30-39	Any	1322*	American	Cancer
10	Jean	30-39	Any	1322*	American	Cancer
11	Ken	30-39	Any	1322*	American	Cancer
12	Lewis	30-39	Any	1322*	American	Cancer

Specification of Adversarial Knowledge

- Difficulty: the data publisher does not know precisely what knowledge an adversary has.
- Martin et al. proposed the use of a ***language*** for expressing such knowledge: **quantify the *amount* of knowledge** an adversary could have, and to release data that are resilient to a certain amount of knowledge regardless of the specific content of this knowledge.
- *Lbasic*(*k*) to be the set of all possible conjunctions of *k* implications :

$\mathcal{L}_{basic}(2)$ is

[((Fox has Flu) and (Igor has Cancer)) implies (Ken has Cancer)]

and

[(Helen has Flu) implies ((Fox has Flu) and (Lewis has Cancer))]

(c, k)-safety

- Given knowledge threshold $k > 0$ and confidence threshold $c \in [0, 1]$, release candidate D^* is (c,k)-safety if

$$\max_{t \in T, s \in S, K \in \mathcal{L}_{basic}(k)} Pr(t \ has \ s | K, D^*) < c$$

where T is the set of individuals involved in D and S is the set of sensitive attribute values.

- Problem: Lbasic(k) is not intuitive, it hard to set an appropriate k value in practice.

3D privacy criterion

» Quantify possible adversarial knowledge from three intuitive dimensions.

» Language $L_{t,s}(l, k, m)$ to be the set of all logic sentences, each of which represents an adversary that knows:

- (1) $\boldsymbol{l}$ sensitive values that the target individual $\boldsymbol{t}$ **does not have**,
- (2) the sensitive values of $\boldsymbol{k}$ **other individuals**,
- (3) $\boldsymbol{m}$ **individuals in** $\boldsymbol{t}$**'s *same-value family*** for a sensitive value *s* (meaning that we can be sure that *t* has sensitive value *s* if any one of those *m* individuals has *s*, especially if *s* is a contagious disease).

3D privacy criterion

- Given knowledge threshold (l, k, m) and confidence threshold $c \in [0,1]$, release candidate D^* is safe if

$$\max_{t \in T, s \in S, K \in \mathcal{L}_{t,s}(l,k,m)} Pr(t \text{ has } s | K, D^*) < c,$$

- What is the relation with k-anonymity and l-diversity?

3D privacy criterion

- *3D privacy* → *k*-anonymity
  - identities are considered to be the sensitive values, the knowledge threshold is ($l = ?, k = ?, m = ?$) and the confidence threshold $c = 1$, for all sensitive values.

- 3D privacy → (c,l)-diversity
  - the knowledge threshold is ($l = ?, k = ?, m = ?$)and the confidence threshold is ?, for all sensitive values.
  - (c, l)-diversity: $r_1 < c(r_l + r_{l+1} + \ldots + r_m)$.

3D privacy criterion

- *3D privacy → k-anonymity*
  - identities are considered to be the sensitive values, the knowledge threshold is $(0, k-2, 0)$ and the confidence threshold $c = 1$, for all sensitive values.
  - k-anonymity provides privacy protection against any adversarial knowledge about the identities of k − 2 individuals.

- 3D privacy → (c,l)-diversity
  - the knowledge threshold is $(l-2, 0, 0)$ and the confidence threshold is $c/(c+1)$, for all sensitive values.
  - (c, l)-diversity: $r_1 < c(r_l + r_{l+1} + \ldots + r_m)$. The recursive (c, l)-diversity guards against all adversaries who possess at most $l-2$ statements of the form “Bob does not have heart disease”

Computation of Breach Probabilities

- Privacy breach: $\Pr(t\ has\ s|K, D^*)$is generally computed based on the random world assumption.
  - Given a release candidate D^*, each possible original data D that can produce D^* by applying the sanitization mechanism to D is called a possible world of D^*, and each possible world is equally likely.

$$Pr(t\ has\ s|K, D^*) = \frac{n((t\ has\ s)\ and\ K|D^*)}{n(K|D^*)}$$

- Compute max Pr: analyze the necessary conditions of the maximum, dynamic programming.

Probabilistic Background Knowledge

- Adversaries may possess **probabilistic knowledge** about parts of the domain.

- The adversary's background knowledge is captured in terms of the prior probability, and additional information due to the access to T^* represented by the posterior probability.

(α,β)-privacy

» Let R be an algorithm that takes as input $u \in D_U$ and outputs $v \in D_V$. R is said to allow an

- **upward (α,β)-privacy breach** with respect to a predicate ∅ if for some probability distribution f,

$$\exists u \in D_U, \exists v \in D_V \ s.t. \ P_f(\Phi(u)) \leq \alpha \ and \ P_f(\Phi(u)|R(u) = v) \geq \beta$$

- **downward (α,β)-privacy breach** with respect to a predicate ∅ if for some probability distribution f,

$$\exists u \in D_U, \exists v \in D_V \ s.t. \ P_f(\Phi(u)) \geq \alpha \ and \ P_f(\Phi(u)|R(u) = v) \leq \beta$$

- R is said to satisfy **(α,β)-privacy** if it does not allow any (α,β)-privacy breach for any predicate ∅.

Positive & Negative Disclosure

- **Positive Disclosure**: Publishing the table $T^\star$ that was derived from T results in a positive disclosure if the adversary can **correctly identify** the value of a sensitive attribute with high probability

- **Negative disclosure**: Publishing the table $T^\star$ that was derived from T results in a negative disclosure if the adversary can **correctly eliminate** some possible values of the sensitive attribute with high probability

(α,β)-privacy

- Unlike previous privacy criteria which define whether or not a release candidate is safe, **(α,β)-privacy defines whether an anonymization algorithm is safe**.

- (α,β)-privacy **considers all possible inputs** (no matter what the data publisher's original data set is) and **all possible outputs** (no matter what release candidate is actually published) of an anonymization algorithm.

- Difficulty: There are far too many possible properties to check them all; we do not know the prior distribution P_f.

(α,β)-privacy

- Idea: compare the operator's transitional probabilities $P(u \to v)$ for the same v but different u. Intuitively, if all of the u-values are reasonably likely to be mapped into a given v, then revealing $R(u) = v$ does not tell too much about u.
- The necessary and sufficient conditions for R to satisfy (α,β)-privacy *for any prior distribution and any property φ*
  - ***γ-amplifying:***

$$\forall v \in D_V, \forall u_1, u_2 \in D_U, \frac{P(R(u_1) = v)}{P(R(u_2) = v)} \leq \gamma$$

  - *Let R be an algorithm that is γ-amplifying. R does not permit an (α,β)-privacy breach for any adversarial prior distribution if and only if*

$$\gamma \leq \frac{\beta}{\alpha} \frac{1-\alpha}{1-\beta}$$

(α,β)-privacy

- (α,β) condition **does not limit the information known to the adversary** as it considers every possible adversarial prior belief.
- If R deterministically maps two inputs u_1 and u_2 to two distinct outputs v_1 and v_2, its amplification is ?

$$\frac{P(R(u_1)) = v_1}{P(R(u_2)) = v_1} = \quad ?$$

(α,β)-privacy

- (α,β) condition **does not limit the information known to the adversary** as it considers every possible adversarial prior belief.
- If R deterministically maps two inputs u_1 and u_2 to two distinct outputs v_1 and v_2, its amplification is ?

$$\frac{P(R(u_1)) = v_1}{P(R(u_2)) = v_1} = \frac{P(R(u_1) = v_1)}{0} = \infty$$

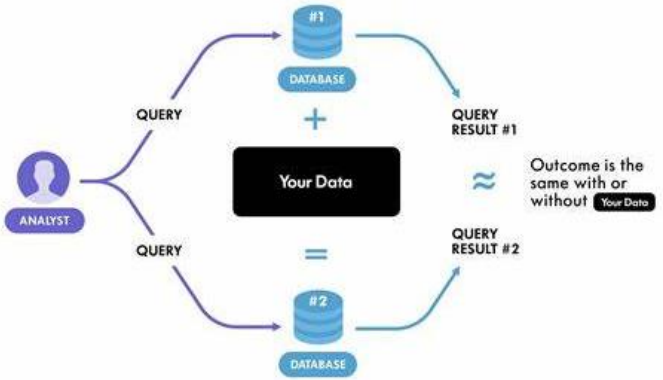
- No deterministic algorithm (e.g., generalization schemes) can satisfy (α,β)-privacy, unless R maps all the inputs to the same output.

Exercise

- Suppose that private information x is a number between 0 and 1000. This number is chosen as a random variable X such that 0 is 1%-likely whereas any non-zero is only about 0.1%-likely.
- Suppose we want to randomize such a number by replacing it with a new random number $y = R(x)$ that retains some information about the original number x. Here are three possible ways to do it:
  - 1. Given x, let $R_{1(x)}$ be x with 20% probability, and some other number (chosen uniformly at random) with 80% probability.
  - 2. Given x, let $R_{2(x)}$ be $x + \delta (mod\ 1001)$, where δ is chosen uniformly at random in{-100, ..., 100}.
  - 3. Given x, let $R_{3(x)}$ be $R_{2(x)}$ with 50% probability, and a uniformly random number otherwise.
  - Compute prior and posterior probabilities of two properties of x: 1) X=0; 2) $X \notin \{200, \dots, 800\}$. Do they satisfy the amplification condition?

Differential Privacy

- The differential privacy criterion, proposed by Dwork (2006), is designed to guarantee the privacy of individuals.



- Similar to (α,β)-privacy, differential privacy defines whether or not **an anonymization algorithm is safe over all possible inputs and outputs.**

Differential Privacy

- Even if the participant removed her data from the data set, no outputs (and thus consequences of outputs) would become significantly more or less likely.

Definition 2. *A randomized function $\mathcal{K}$ gives* ϵ-differential privacy *if for all data sets D_1 and D_2 differing on at most one element, and all $S \subseteq Range(\mathcal{K})$,*

$$\Pr[\mathcal{K}(D_1) \in S] \leq \exp(\epsilon) \times \Pr[\mathcal{K}(D_2) \in S] \quad (1)$$

- Semantically, ε-differential privacy is stronger than (α,β)-privacy, since the latter only considers adversarial knowledge about a single individual, but the former considers adversarial knowledge about all individuals in the table.
- However, by adding adversarial knowledge of exact information about “all but one” individuals in the table, we showed that the variant (α, εα)-individual privacy (for all α) is equivalent to ε-differential privacy.

Perfect Privacy

- Some data is so secret that an individual may not want *any* information to be disclosed. Such a stringent privacy requirement is termed ***perfect privacy*** and is equivalent to Shannon's notion of perfect secrecy.

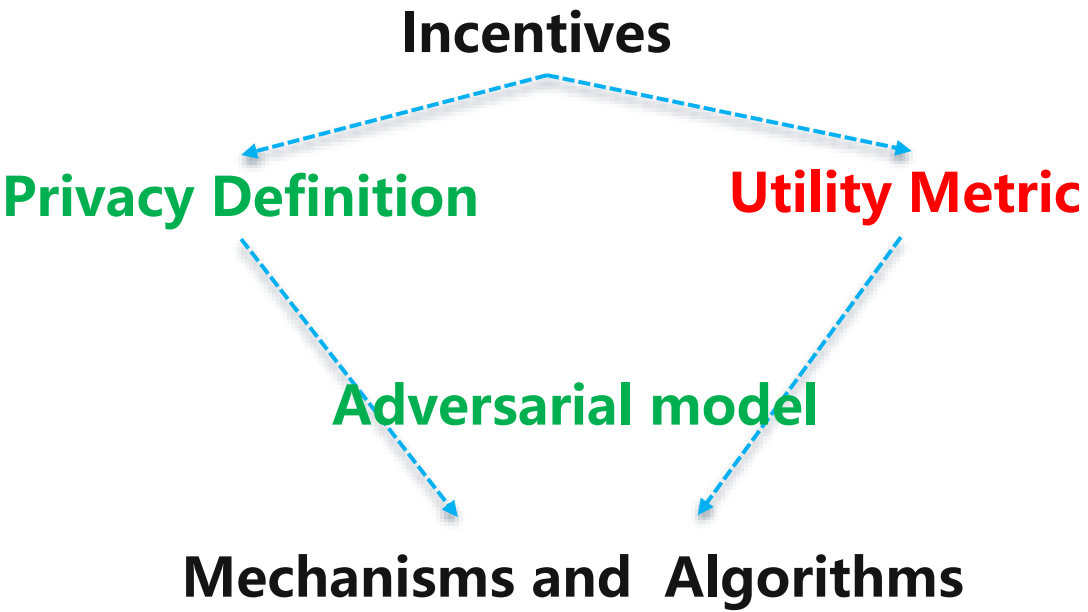
- Publishing a view of a relational table *T* violates perfect privacy if belief about the answer to the secret query **changes** after seeing the published view.

$$P_f(Q_S(T) = S) \neq P_f(Q_S(T) = S \,|\, Q_V(T) = V)$$

Dicussion

- “Which privacy definition should be used for a specific application?”
- Unfortunately, there is neither a mandate on how to define privacy for a new application, nor a clear technique to compare the various privacy definitions prevalent in the literature.
- The problem can be solved if all privacy definitions can be expressed under one common framework.

How Much


Incentives
Privacy Definition
Utility Metric
Adversarial model
Mechanisms and Algorithms

5. Measures of Utility

A data publisher seeks to release data that are not only safe, but also useful.

Measures of Utility

- Quantitative measures of information loss
  - Simple example: number of rows suppressed in a table
- Evaluate the quality of the results for all queries in some fixed class
  - Hope the class is representative, so other uses have low distortion
  - Costly: some methods enumerate all queries, or all anonymizations
- Empirical Evaluation
  - Perform experiments with a reasonable workload on the result
  - Compare to results on original data (e.g. Netflix prize problems)
- Combinations of multiple methods
  - Optimize for some surrogate, but also evaluate on real queries

Generalization/Suppression Counting

- **number of anonymization operations** performed on a data set.
  - E.g., if generalization is the only operation being performed, then it is reasonable to measure information loss by the number of generalization steps performed.
- Samarati used one version called generalization height.
- Meyerson and Williams used another variation: they measured the total number of attribute values that were suppressed.
- Problem: not all operations affect utility in the same way.

Loss Metric (LM)

- LM is defined in terms of a normalized loss for each attribute of every tuple.

- For a tuple t and categorical attribute A, suppose the value of t[A] has been generalized to x. Letting |A| represent the size of the domain of attribute A and letting M represent the number of values in this domain that could have been generalized to x, then the **loss for t[A] is (M − 1)/(|A| − 1).**

- The loss for attribute A is defined as the average of the loss t[A] for all tuples t. The LM for the entire data set is defined as the sum of the losses for each attribute.

Classification Metric (CM)

- The classification metric (CM) is designed to measure the **effect of the anonymization on a hypothetical classifier**.

- In this scenario, there is a distinguished class attribute, and tuples are placed into groups (by quasi-identifier value). Each tuple incurs a penalty of 1 if it is suppressed or if its class attribute is not the same as the majority class attribute in the group.

- The classification metric is defined as the average of the penalties of all the tuples.

Discernibility Metric (DM)

- A metric similar in spirit to LM.
- DM assigns a penalty to each tuple based on how many other tuples in the database are indistinguishable from it, and therefore it works naturally in the k-anonymity framework.
- For a database of size n, DM assigns a penalty of n for each suppressed tuple. If a tuple is not suppressed, the penalty it receives is the total number of tuples in the database having the same quasi-identifier values.

Ambiguity Metric (AM)

- Nergiz and Clifton proposed *ambiguity metric* (AM), that is especially suitable for the *k*-anonymity framework.

- For each tuple $t*$ in the sanitized data, AM considers **the number of tuples in the *domain* of the data that could have been mapped (generalized) to $t*$.** This number is the ambiguity of $t*$.

- The AM for the sanitized data set is then the average ambiguity for all tuples in the sanitized data.

KL-divergence /LP-Norm/Hellinger Distance

- If age was uniformly distributed, and independent of all other attributes, then replacing the age attribute with an age range would have little effect.
- Measure the distance between the original probability distribution and the probability distribution reconstructed from the sanitized data.
- The larger the distance is, the greater the information loss.

Reconstructibility

- The approaches discussed so far measure the utility of the sanitized data that are actually produced.
- Measure utility in terms of the algorithm used to create the sanitized data, in this case, the result is usually a probabilistic utility guarantee.
- Eg., Agrawal et al. define the utility associated with a randomized anonymization algorithm in terms of the ability to reconstruct statistics from the sanitized data.

- If f is a real-valued function computed over the original data, and f' is the estimator of f computed over the sanitized data, then f is (n, ε, δ) reconstructible if $|f - f'| < \max(\varepsilon, \varepsilon f)$ with probability at least $(1 - \delta)$ whenever the number of tuples in the original data is at least n.

Other Metrics

- Invariance
- Analytical Validity
- First- and Second-Order Statistics
- Workload-Aware Metrics
- Bivariate Measures

THANKS!

Any questions?

You can find me at:

- zhanglan@ustc.edu.cn