PRIVACY PRESERVING DATA PUBLISHING

Digital security and privacy

Georgios Spathoulas

Msc in ”Informatics and computational bio-medicine”

University of Thessaly

INTRODUCTION

DATA PUBLISHING

- The collection of digital information by governments, corporations, and individuals has created tremendous opportunities for knowledge-based decision making
- For example, licensed hospitals in California are required to submit specific demographic data on every patient discharged from their facility
- Data publishing is equally ubiquitous in other domains
    - For example, **Netflix**, a popular online movie rental service, recently published a data set containing movie ratings of 500,000 subscribers, in a drive to improve the accuracy of movie recommendations based on personal preferences (New York Times, Oct. 2, 2006);
    - **AOL** published a release of query logs but quickly removed it due to the reidentification of a searcher [Barbaro and Zeller 2006].

PRIVACY PRESERVING DATA PUBLISHING

- **Detailed person-specific data** in its original form often contains **sensitive information about individuals**, and publishing such data immediately **violates individual privacy**
- The current practice primarily relies on policies and guidelines to **restrict the types of publishable data** and on agreements on the **use and storage of sensitive data**
- The **limitation** of this approach is that it either **distorts data excessively** or requires a **trust level** that is **impractically high** in many data-sharing scenarios.
- A task of the utmost importance is to develop methods and tools for publishing data in a more hostile environment, so that the **published data remains practically useful** while **individual privacy is preserved**
- This undertaking is called **privacy-preserving data publishing (PPDP)**

Privacy-preserving data publishing

DATA COLLECTION AND PUBLICATION PHASES

- In the **data collection phase**, the data publisher collects data from record owners (e.g., Alice and Bob)
- In the **data publishing phase**, the data publisher releases the collected data to a data miner or to the public, called the data recipient, who will then conduct data mining on the published data
- For example, a **hospital** collects **data** from **patients** and **publishes** the patient records to an **external medical center**
- In this example, the **hospital** is **the data publisher**, **patients** are **record owners**, and the **medical center** is the **data recipient**
- The data mining conducted at the medical center could be anything from a simple count of the number of men with diabetes to a sophisticated cluster analysis

MODELS OF DATA PUBLISHERS

There are two models of data publishers :

- In the **untrusted model**, the **data publisher is not trusted** and may attempt to identify sensitive information from record owners
- Various cryptographic solutions, anonymous communications and statistical methods were proposed to collect records anonymously from their owners without revealing the owners' identity
- In the **trusted model**, the **data publisher is trustworthy** and record owners are willing to provide their personal information to the data publisher; however, the trust **is not transitive to the data recipient**
- We assume the trusted model of data publishers and consider privacy issues in the data publishing phase

DATA FORMAT

In the basic form of PPDP, the data publisher has a table of the form

D(Explicit Identifier, Quasi Identifier, Sensitive Attributes, Non-Sensitive Attributes), where :

- **Explicit Identifier** is a set of attributes, such as name and social security number, containing information that explicitly identifies record owners
- **Quasi Identifier (QID)** is a set of attributes that could potentially identify record owners
- **Sensitive Attributes** consists of sensitive person-specific information such as disease, salary, and disability status
- and **Non-Sensitive Attributes** contains all attributes that do not fall into the previous three categories

ANONYMIZATION

- **Anonymization** refers to the PPDP approach that seeks to hide the identity and/or the sensitive data of record owners, assuming that sensitive data must be retained for data analysis
- Clearly, **explicit identifiers** of record owners must be removed
- Even with all explicit identifiers being removed, a past research showed a **real-life privacy threat to William Weld**, former governor of the state of Massachusetts
- An individual's name in a **public voter list** was **linked** with his **record** in a **published medical database** through the combination of **zip code, date of birth, and sex**
- Each of these attributes does not uniquely identify a record owner, but their combination, called the **quasi-identifier**, often singles out a unique or a small number of record owners
- **87%** of the U.S. population had reported characteristics that likely made them unique based on only such quasi-identifiers

LINKING ATTACKS

- In the above example, the owner of a record is re-identified by **linking** his **quasi-identifier**
- To perform such linking attacks, the attacker needs two pieces of prior knowledge:
  - the victim's record in the released data
  - and the quasi-identifier of the victim
- Such knowledge can be obtained by observation
- For example, the attacker noticed that his boss was hospitalized, and therefore knew that his boss's medical record would appear in the released patient database
- Also, it was not difficult for the attacker to obtain his boss's zip code, date of birth, and sex, which could serve as the quasi-identifier in linking attacks

PREVENTING LINKING ATTACKS

To prevent linking attacks, the data publisher provides an anonymous table:

T (QID' , Sensitive Attributes, Non-Sensitive Attributes)

- **QID'** is an anonymous version of the original **QID** obtained by applying anonymization operations to the attributes in **QID** in the original table D
- Anonymization operations hide some detailed information so that several records become indistinguishable with respect to **QID**
- Consequently, if a person is linked to a record through **QID**, that person is also linked to all other records that have the same value for **QID** , making the linking ambiguous
- The anonymization problem is to produce an anonymous T that satisfies a given privacy requirement determined by the chosen privacy model and to retain as much data utility as possible

Attack models and privacy models

PRIVACY PROTECTION

Access to the published data **should not enable the attacker to learn anything extra** about any target victim **compared to no access to the database**, even with the presence of any attacker's background knowledge obtained from other sources

Privacy models can be classified into two categories based on their attack principles:

- **Linkage attacks:** An attacker is able to link a record owner
  - to a record in a published data table **(record linkage)**
  - to a sensitive attribute in a published data table **(attribute linkage)**
  - to the published data table itself **(table linkage)**
- **Probabilistic attacks:** If the attacker **has a large variation between the prior and posterior beliefs**

RECORD LINKAGE

- In the attack of record linkage, **some value qid on QID identifies a small number of records** in the released table T , called a group
- If the victim's QID matches the value qid, the victim is vulnerable to being linked to the small number of records in the group
- In this case, the attacker faces only a small number of possibilities for the victim's record, and with the help of additional knowledge, there is a chance that the attacker could uniquely identify the victim's record from the group

EXAMPLES

(a) Patient table

Job	Sex	Age	Disease
Engineer	Male	35	Hepatitis
Engineer	Male	38	Hepatitis
Lawyer	Male	38	HIV
Writer	Female	30	Flu
Writer	Female	30	HIV
Dancer	Female	30	HIV
Dancer	Female	30	HIV

(b) External table

Name	Job	Sex	Age
Alice	Writer	Female	30
Bob	Engineer	Male	35
Cathy	Writer	Female	30
Doug	Lawyer	Male	38
Emily	Dancer	Female	30
Fred	Engineer	Male	38
Gladys	Dancer	Female	30
Henry	Lawyer	Male	39
Irene	Dancer	Female	32

(c) 3-anonymous patient table

Job	Sex	Age	Disease
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	HIV
Artist	Female	[30-35)	Flu
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV

(d) 4-anonymous external table

Name	Job	Sex	Age
Alice	Artist	Female	[30-35)
Bob	Professional	Male	[35-40)
Cathy	Artist	Female	[30-35)
Doug	Professional	Male	[35-40)
Emily	Artist	Female	[30-35)
Fred	Professional	Male	[35-40)
Gladys	Artist	Female	[30-35)
Henry	Professional	Male	[35-40)
Irene	Artist	Female	[30-35)

Examp

Illustrating Various Attacks

EXAMPLE

- Suppose that a hospital wants to publish the patient records in **Table (a)** to a research center
- Suppose that the research center has access to the external table **Table (b)** and knows that every person with a record in Table (b) has a record in Table (a)
- **Joining** the two tables **on the common attributes** Job, Sex, and Age may link the identity of a person to his/her Disease
- For example, **Doug**, a male lawyer who is 38 years old, is identified as an **HIV patient** by qid = Lawyer, Male, 38 after the join
- What about **Alice**?

K-ANONYMITY

To prevent record linkage through QID, the notion of **k-anonymity** has been proposed:

k-anonymity

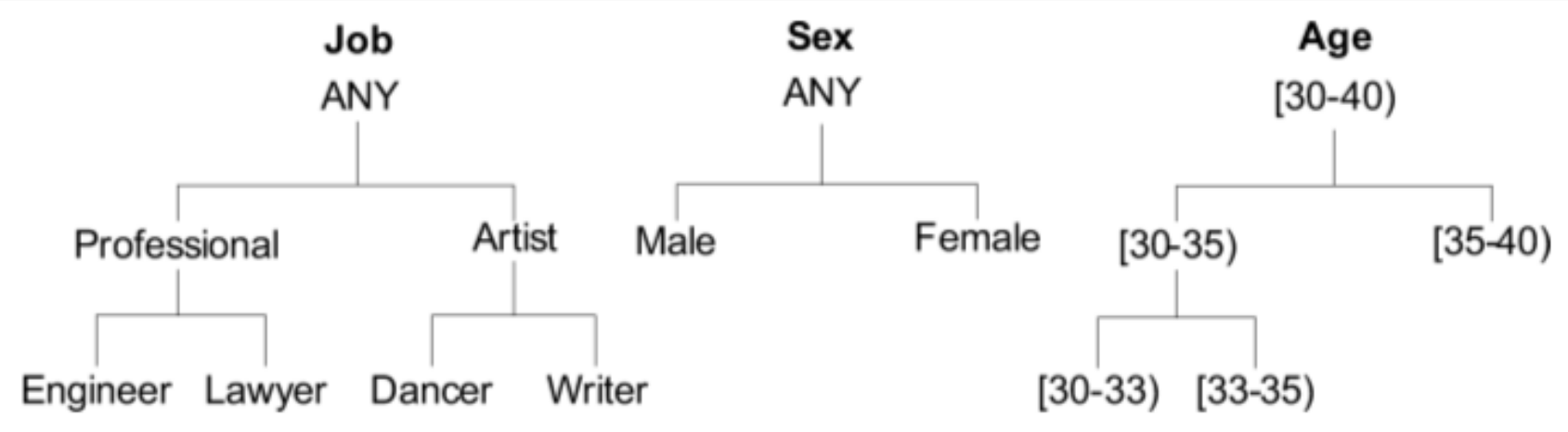
if one record in the table has some value **qid**, at least **k − 1** other records also have the same value **qid**

- In other words, the minimum group size on QID is at least k
- A table satisfying this requirement is called k-anonymous
- In a k-anonymous table, each record is indistinguishable from at least k − 1 other records with respect to QID
- Consequently, the probability of linking a victim to a specific record through QID is at most 1/k

EXAMPLE

- Table (c) shows a **3-anonymous table** by generalizing QID = Job, Sex, Age from Table (a) using the taxonomy trees on next slide
- It has two distinct groups on QID, namely
  - Professional, Male, [35–40)
  - and Artist, Female, [30–35)
- Since each group contains at least 3 records, **the table is 3-anonymous**
- If we link the records in Table (b) to the records in Table (c) through QID, each record is linked to either no record or at least 3 records in Table (c)

TAXONOMY TREES



Taxonomy trees

MULTIPLE QIDS

- The k-anonymity model assumes that QID is known to the data publisher
- Most work considers a single QID containing all attributes that can be potentially used in the quasi-identifier
- The more attributes included in QID, the more protection k-anonymity would provide
- On the other hand, this also implies that more distortion is needed to achieve k-anonymity because the records in a group have to agree on more attributes
- To address this issue, **multiple QIDs** can be specified assuming that the data publisher knows the potential QIDs for record linkage

MULTIPLE QIDS

- The data publisher wants to publish a table **T (A, B, C, D, S)**, where S is the sensitive attribute, and knows that the data recipient has access to previously published tables **T_1(A, B, X)** and **T_2(C, D, Y)**, where X and Y are attributes not in T
- To prevent linking the records in T to the information on X or Y , the data publisher can specify **k-anonymity** on **QID_1 = A, B** and **QID_2 = C, D** for T
- This means that each record in T is indistinguishable from a group of at least k records with respect to QID_1 and is indistinguishable from a group of at least k records with respect to QID_2
- The two groups are not necessarily the same
- Clearly, this requirement is implied by k-anonymity on QID = A, B, C, D, but having k-anonymity on both QID_1 and QID_2 does not imply k-anonymity on QID

ATTRIBUTE LINKAGE

- In the attack of **attribute linkage**, the attacker may not precisely identify the record of the target victim, but could infer his/her sensitive values from the published data T, based on the set of sensitive values associated to the group that the victim belongs to
- In case some sensitive values predominate in a group, **a successful inference becomes relatively easy even if k-anonymity is satisfied**.
- Several other approaches have been proposed to address this type of threat
- The general idea is to diminish the correlation between QID attributes and sensitive attributes

ATTRIBUTE LINKAGE EXAMPLE

- From **Table (a)**, an attacker can infer that all female dancers at age 30 have HIV
- **Dancer, Female, 30 → HIV with 100% confidence**
- Applying this knowledge to **Table (b)**, the attacker can infer that Emily has HIV with 100% confidence provided that Emily comes from the same population in **Table (a)**

ℓ-DIVERSITY

- The diversity principle, called **ℓ-diversity** has been proposed to prevent attribute linkage
- The ℓ-diversity requires every qid group to contain at least ℓ "well-represented" sensitive values
- The simplest understanding of "well-represented" is to ensure that **there are at least ℓ distinct values for the sensitive attribute in each qid group**
- This distinct ℓ-diversity privacy model automatically satisfies k-anonymity, where k = ℓ, because each qid group contains at least ℓ records
- Some **sensitive values** are naturally **more frequent** than others in a group, enabling an attacker to conclude that a record in the group is very likely to have those values. For example, Flu is more common than HIV.
- This motivates a stronger notion of ℓ-diversity, **entropy ℓ-diversity**

ENTROPY ℓ-DIVERSITY

A table is entropy ℓ-diverse if for every qid group :

$$-\sum_{S} P(qid, s) \log(P(qid, s)) \geq \log(\ell) \tag{1}$$

- where **S** is a **sensitive attribute**, and **P(qid, s)** is the **fraction of records in a qid group having the sensitive value s**
- The left-hand side, called the entropy of the sensitive attribute, has the property that more evenly distributed sensitive values in a qid group produce a larger value
- Therefore, a large threshold value implies less certainty of inferring a particular sensitive value in a group

ENTROPY ℓ-DIVERSITY EXAMPLE

- Consider Table (c)
- For the first group Professional, Male, [35 -40) ,
  $-\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} = \log(1.9)$
- and for the second group Artist, Female, [30-35) ,
  $-\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} = \log(1.8)$
- So the table satisfies entropy ℓ-diversity if $\ell \leq 1.8$

ANONYMIZATION OPERATIONS

ANONYMIZATION OPERATIONS

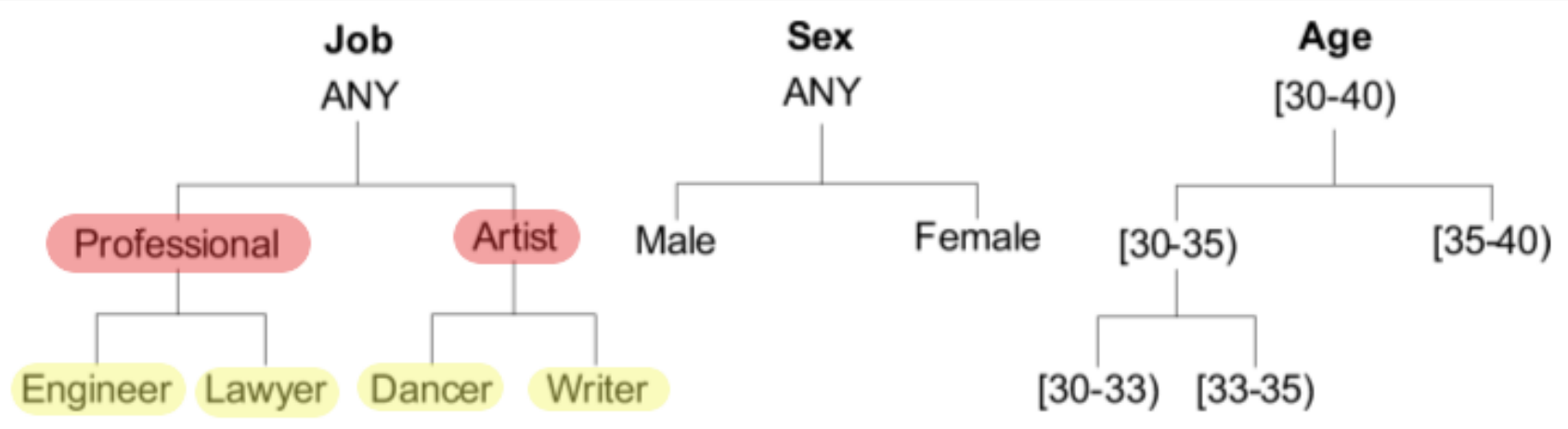
- Typically, the original table does not satisfy a specified privacy requirement
- The modification is done by applying a sequence of anonymization operations to the table
- Anonymization comes in several flavors: **generalization, suppression, anatomization, permutation**, and **perturbation**
- **Generalization** and **suppression** replace values of specific description, typically the QID attributes, with less specific description
- **Anatomization** and **permutation** deassociate the correlation between QID and sensitive attributes by grouping and shuffling sensitive values in a qid group
- **Perturbation** distorts the data by adding noise, aggregating values, swapping values, or generating synthetic data based on some statistical properties of the original data

GENERALIZATION AND SUPPRESSION

- Each generalization or suppression operation hides some details in QID
- For a categorical attribute, a specific value can be replaced with a general value according to a given taxonomy
- For a numerical attribute, exact values can be replaced with an interval that covers exact values
- If a taxonomy of intervals is given, the situation is similar to categorical attributes
- A **generalization** replaces some values with **a parent value** in the taxonomy of an attribute
- A **suppression** replaces some values with **a special value**, indicating that the replaced values are not disclosed

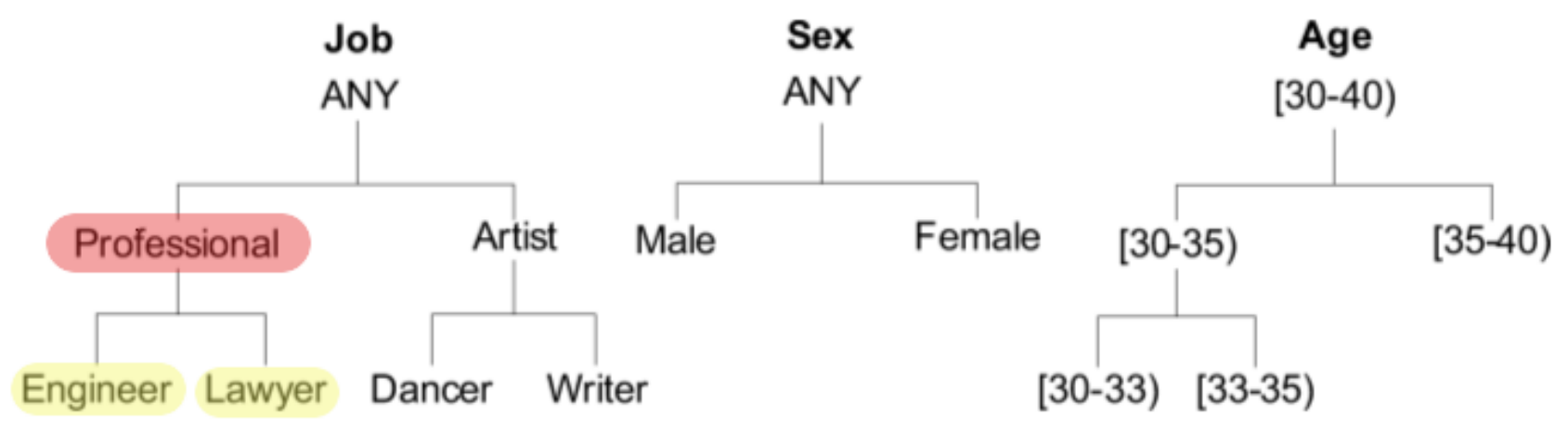
FULL-DOMAIN GENERALIZATION SCHEME

- In this scheme, all values in an attribute are generalized **to the same level of the taxonomy tree**
- For example, if Lawyer and Engineer are generalized to Professional, then it also requires generalizing Dancer and Writer to Artist
- The search space for this scheme is much smaller than the search space for other schemes below, but the data distortion is the largest because of the same granularity level requirement on all paths of a taxonomy tree



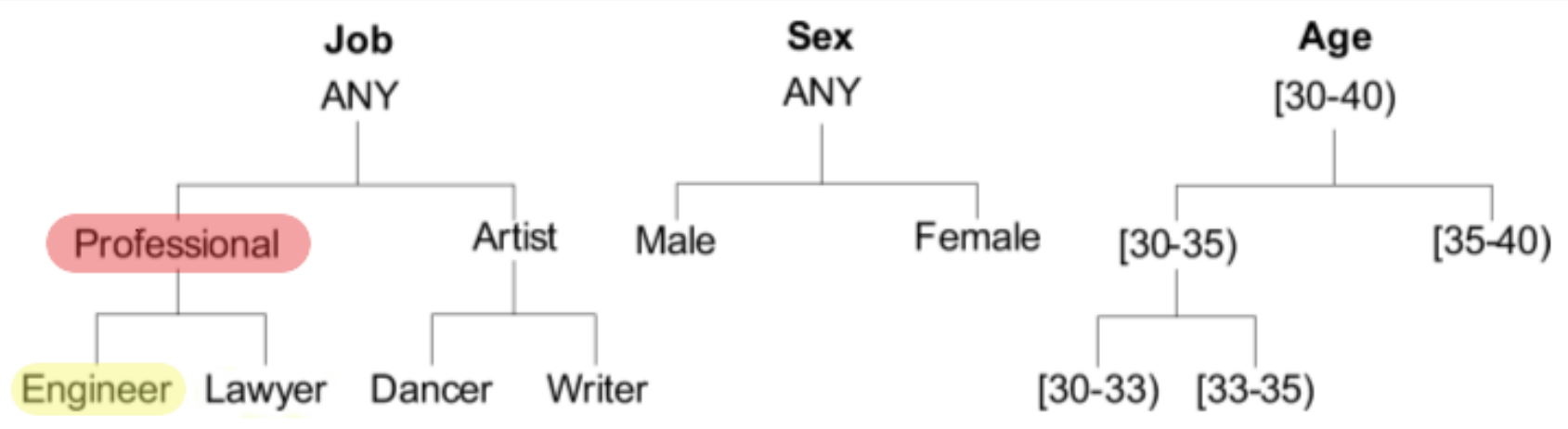
SUBTREE GENERALIZATION SCHEME

- In this scheme, at a nonleaf node, either **all child values** or **none** are generalized
- For example, if Engineer is generalized to Professional, this scheme also requires the other child node, Lawyer, to be generalized to Professional, but Dancer and Writer, which are child nodes of Artist, can remain ungeneralized



SIBLING GENERALIZATION SCHEME

- This scheme is similar to the sub-tree generalization, except that **some siblings may remain ungeneralized**
- A parent value is then interpreted as representing all missing child values
- For example, if Engineer is generalized to Professional, and Lawyer remains ungeneralized, Professional is interpreted as all jobs covered by Professional except for Lawyer
- This scheme produces less distortion than sub-tree generalization schemes because it only needs to generalize the child nodes that violate the specified threshold



MEASURING INFORMATION LOSS

ILoss is a data metric to capture the information loss of generalizing a specific value to a general value

If, for an **attribute A**, a **specific value v** is generalized to a **more general value v_g** then the information loss is computed as the fraction of values being generalised to v_g, to the whole values of attribute A

$$\text{ILoss}(u_g) = \frac{|u_g| - 1}{|D_A|} \tag{2}$$

- $|u_g|$: Population of values of A being generalized
- $|D_A|$: General population of A values

MEASURING INFORMATION LOSS

In order to calculate the information loss for **a whole record** its attribute losses are summed with a use of a constant weight for each attribute

$$\mathrm{ILoss}(r) = \sum_{u_g \in r} w_i * \mathrm{ILoss}(u_g) \tag{3}$$

In order to calculate the information loss for a **whole generalised table** its records losses are summed up

$$\mathrm{ILoss}(T) = \sum_{r \in T} \mathrm{ILoss}(r) \tag{4}$$

ANATOMIZATION

- Unlike generalization, **anatomization** does not modify the quasi-identifier or the sensitive attribute, but **deassociates the relationship** between the two
- Precisely, the method releases the data on QID and the data on the sensitive attribute in **two separate tables**: a **quasi-identifier table (QIT)** contains the QID attributes, a **sensitive table (ST)** contains the sensitive attributes, and both QIT and ST have **one common attribute, GroupID**
- All records in the same group will have the same value on GroupID in both tables, and therefore are linked to the sensitive values in the group in the exact same way
- If a group has ℓ distinct sensitive values and each distinct value occurs exactly once in the group, then the probability of linking a record to a sensitive value by GroupID is $\frac{1}{\ell}$
- The attribute linkage attack can be distorted by increasing ℓ.

ANATOMIZATION EXAMPLE

- Suppose that the data publisher wants to release the patient data in Table (a), where **Disease is a sensitive attribute** and **QID = {Age, Sex}**
- First, partition (or generalize) the original records into qid groups so that, in each group, **at most $\frac{1}{\ell}$ of the records contain the same Disease value**
- This intermediate Table (b) contains two qid groups: [30–35), Male and [35–40), Female
- Next, create **QIT (Table (c))** to contain all records from the original **Table (a)**, but **replace the sensitive values by the GroupIDs**, and create **ST (Table (d))** to contain the **count of each Disease for each qid group**
- QIT and ST satisfy the privacy requirement with $\ell \leq 2$ because each qid group in QIT infers any associated Disease in ST with probability at most $\frac{1}{\ell} = \frac{1}{2} = 50\%$

ANATOMIZATION EXAMPLE

(a) Original patient data

Age	Sex	Disease (sensitive)
30	Male	Hepatitis
30	Male	Hepatitis
30	Male	HIV
32	Male	Hepatitis
32	Male	HIV
32	Male	HIV
36	Female	Flu
38	Female	Flu
38	Female	Heart
38	Female	Heart

(b) Intermediate *QID*-grouped table

Age	Sex	Disease (sensitive)
[30–35)	Male	Hepatitis
[30–35)	Male	Hepatitis
[30–35)	Male	HIV
[30–35)	Male	Hepatitis
[30–35)	Male	HIV
[30–35)	Male	HIV
[35–40)	Female	Flu
[35–40)	Female	Flu
[35–40)	Female	Heart
[35–40)	Female	Heart

(c) Quasi-identifier table (QIT) for release

Age	Sex	GroupID
30	Male	1
30	Male	1
30	Male	1
32	Male	1
32	Male	1
32	Male	1
36	Female	2
38	Female	2
38	Female	2
38	Female	2

(d) Sensitive table (ST) for release

GroupID	Disease (sensitive)	Count
1	Hepatitis	3
1	HIV	3
2	Flu	2
2	Heart	2

ANATOMY ADVANTAGE

- The major advantage of anatomy is that **the data in both QIT and ST is unmodified**
- The anatomized tables can more accurately answer aggregate queries involving domain values of the QID and sensitive attributes than the generalization approach
- The intuition is that, in a generalized table, domain values are lost, and without additional knowledge, the uniform distribution assumption is the best that can be used to answer a query about domain values
- In contrast, all domain values are retained in the anatomized tables, which give the exact distribution of domain values

ANATOMY ADVANTAGE EXAMPLE

- For instance, suppose that the data recipient wants to count the **number of patients of age 38 having heart disease**
- The **correct count** from the original Table (a) is **2**
- The **expected count** from the anatomized Table (c) and Table (d) is $3 * \frac{2}{4} = \mathbf{1.5}$, since 2 out of the 4 records in GroupID = 2 in Table (d) have heart disease
- This count is more accurate than the **expected count** $2 * \frac{1}{5} = \mathbf{0.4}$, from the generalized Table (b), where the $\frac{1}{5}$ comes from the fact that the 2 patients with heart disease have an equal chance to be of age 35, 36, 37, 38, 39

DIFFERENTIAL PRIVACY

USING A COIN

- When we **collect data**, participants must respond to questions they are **not comfortable with**
- An approach to solving this problem is to **add some noise** to the data (the answers)
- If this noise is carefully produced then it is possible to
  - Protect users privacy
  - Do not significantly alter the conclusion made by the data
- The oldest and most simple approach is by **using a coin**

USING A COIN

- Imagine the question **”Have you smoked marijuana?”**
- People may feel uncomfortable with the question
- Let’s take the **coin approach**
- Flip a coin
  - If **head** come up $\rightarrow$ **answer truthfully**
  - If **tails** come up $\rightarrow$ **flip a second coin**
    - If **head** come up $\rightarrow$ **answer yes**
    - If **tails** come up $\rightarrow$ **answer no**
- If there are a lot of participants, then the percentage of people that have smoked marijuana may be induced, without violating each participants privacy

CALCULATING PROBABILITIES

- After conducting the survey (n participants) we count the answers
  - x = counted yes
  - y = counted no
- The counted yes are the sum of the yes answers that have been produced by either:
  - Participants that the **first coin** has produced **head** and have **truthfully answered yes**
  - Participants that the **first coin** has produced **tails** and the **second coin** has produced **head**

$$x = \frac{1}{2} * \text{real\_yes} + \frac{n}{4} \tag{5}$$

- So the real yes answers can be calculated :

$$\text{real\_yes} = 2 * x - \frac{n}{2} \tag{6}$$

WHAT ABOUT THE REAL ANSWERS

- If someone have answered yes then the probability that the **randomized answer is yes** is :

$$\Pr[\text{Response} = \text{Yes} | \text{Truth} = \text{Yes}] = \frac{3}{4} \tag{7}$$

  - Either the first coin comes up head (probability 1/2)
  - or the first comes up tails and the second comes up head (probability 1/4)

- If someone have answered yes then the probability that the **randomized answer is no** is :

$$\Pr[\text{Response} = \text{No} | \text{Truth} = \text{Yes}] = \frac{1}{4} \tag{8}$$

  - first comes up tails and second comes up tails (probability 1/4)

WHAT HAPPENS WITH NUMERIC VALUES ?

- There are 50 people in the room. Imagine the question :
- **What is the average income of people in the room ?**
- An now imagine a second question :
- **What is the average income of people in the room excluding the tutor ?**

This combination of questions probably violates tutors privacy

Being member of a group should not decrease your privacy

PRIVACY COMES FROM NOISE

- There is a statistical function : **the average of the incomes**
- By design it does not reveal the income of each participant
- Imagine there are two datasets D and D', where D' comes from D by just removing one record
- **By using the same mechanism on both D and D', I gain information (or violate privacy) for the missing record**
- There is an approach that can be used to enhance the mechanism
- We should add some **noise** to it
- So instead using the **average() function** we use **K(average) function**, where K adds noise

WHAT DOES NOISE LOOKS LIKE

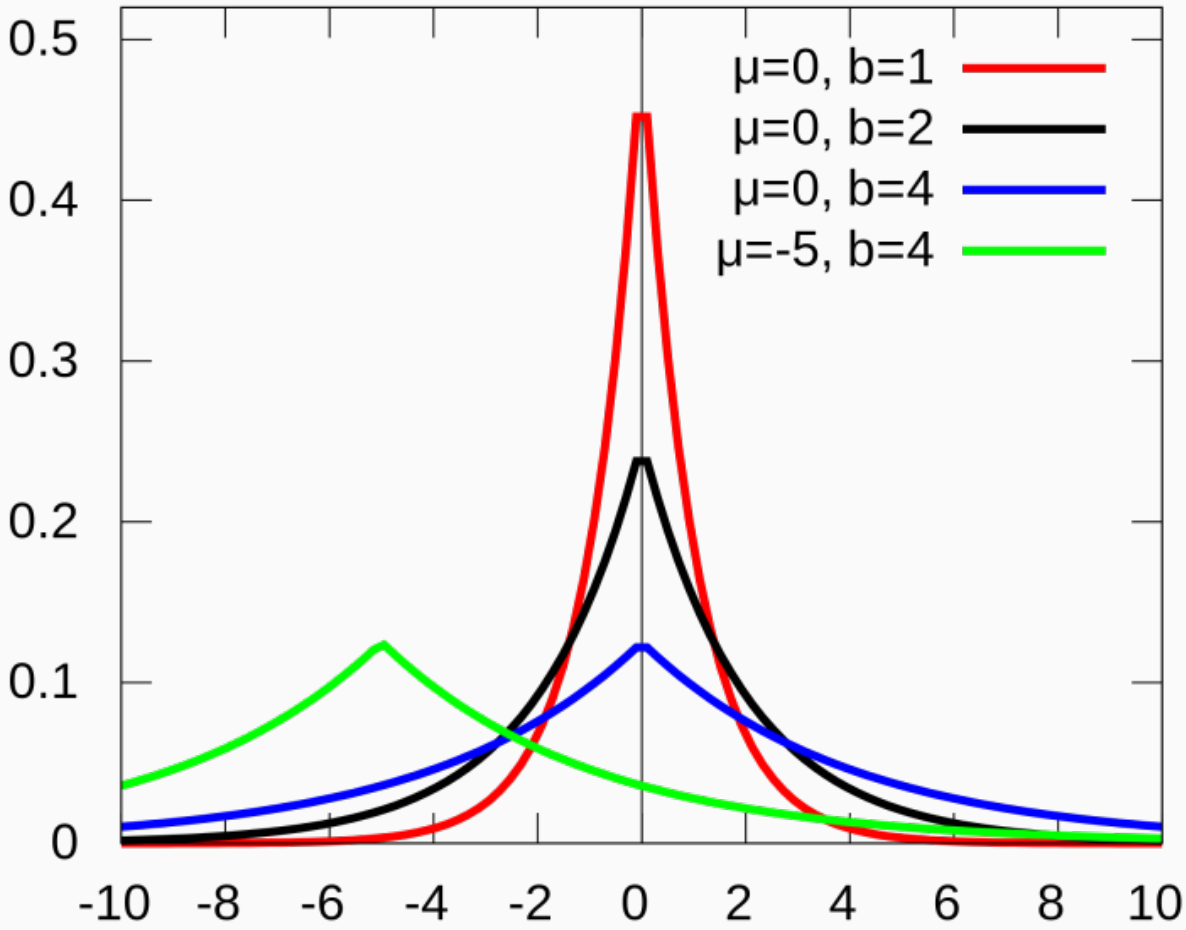


Figure: Laplace distribution

EXAMPLE

- Without noise :
  - Average income of all : 732 €
  - Average income of all without tutor: 728 €
  - Income of tutor = $50 \times 732 - 49 \times 648 = 928$ €
  - We can calculate tutor's income
- With noise :
  - Noise = 3.4 $\rightarrow$ Av. income of all : 732 + 3.4 = 735.4 €
  - Noise = -1.2 $\rightarrow$ Av. income of all without tutor: 728 + (-1.2) = 726.8 €
  - Income of tutor = $50 \times 735.4 - 49 \times 726.8 = 1156.8$ €
  - We make a large mistake in calculating tutor's income, while stats are approximately right

Formally, **differential privacy** is defined as follows:

A **randomized function K** gives **ε-differential privacy** if for all data sets D and D' differing on at most one row, and all $S \subseteq \text{Range}(K)$,

$$\Pr[K(D) \in S] \leq \exp(\varepsilon) \times \Pr[K(D') \in S] \tag{9}$$

This can be translated as meaning that the **risk to one's privacy** should **not substantially** (as bounded by ε) **increase** as a result of **participating** in a statistical database.

LAPLACE MECHANISM

- The **Laplace mechanism** involves adding random noise that conforms to the Laplace statistical distribution
- How much noise should be added ?
- The 0-centered Laplace distribution has only one parameter (its scale)
- It only depends on the privacy parameter, ε and the maximum difference in the values that the query f may take on a pair of databases that differ in only one row

$$\Delta f = \max_{D,D'} ||f(D) - f(D')||_1 \tag{10}$$

- Adding a random Laplace($\Delta f/\varepsilon$) variable to a query, guarantees ε-differential privacy

COMPOSITION

- What happens when we run the query multiple times and take the average...?
- There is a bad property known as **composition**
- For a dataset queried q times, with each query having privacy parameter ε_i, the total privacy budget of the dataset is given by

$$\varepsilon_{\text{total}} = \sum_{i=1}^{q} \varepsilon_i \tag{11}$$

- In practice **ε** is more about a **privacy budget** rather than purely the statistical upper bound of a query
- The **total ε** reflects the **maximum privacy release allowable** for the total query session
- As each query answer leaks privacy, once the budget is exceeded the user will not be able to make any further queries

`http://content.research.neustar.biz/blog/differential-privacy/WhiteQuery.html`

BIBLIOGRAPHY

- Fung, B., Wang, K., Chen, R., & Yu, P. S. (2010). **Privacy-preserving data publishing: A survey of recent developments**. ACM Computing Surveys (CSUR), 42(4), 14.
- Fung, B. C., Wang, K. E., Fu, A. W. C., & Philip, S. Y. (2010). **Introduction to privacy-preserving data publishing: concepts and techniques**. CRC Press.