COMP3260/6360
Data Security
Lecture 11

Prof Ljiljana Brankovic
School of Electrical Engineering and Computer Science

Lecture Overview

- Technical Aspects of Privacy
  - Sample Database
  - Sample Attack
- Types of Attacks
- K-anonymity
- L-diversity
- Differential Privacy

Technical Aspects of Privacy

Resources:

This lecture notes

Note that in-text references and quotes are omitted for clarity of the slides. When you write as essay or a report it is very important that you use both in-text references and quotes where appropriate.

Abstract model

Name	City	Age	Sex	Status	Post-traumatic stress disorder	Attempted suicide
White	Sydney	34	F	W	4.1	no
Scarlet	Dubbo	27	F	D	3.9	no
Brown	Sydney	45	M	M	4.3	no
Mustard	Perth	32	M	S	2.1	yes
Green	Ballina	76	M	M	4.8	no
Green	Darwin	32	F	M	4.6	no
Plum	Hobart	25	M	D	2.9	no
Mustard	Darwin	24	M	W	4.2	no
White	Dubbo	51	F	D	3.8	no
Peacock	Sydney	40	F	M	4.1	no
Black	Ballina	68	F	W	3.6	no
Violet	Dubbo	33	F	M	2.7	no
Aureate	Sydney	28	F	S	3.5	no

Compromise

Example 1

COUNT(City=Darwin and Sex=M and Age<30) = 1

COUNT(City=Darwin and Sex=M and Age<30 and AS=no) = 0

AVG(City=Darwin and Sex=M and Age<30; PTSD) = 4.2

Example 2

COUNT(City=Sydney and Age<37) = 2

COUNT(City=Sydney and Age<37 and AS=no) = 2

Basic Privacy Techniques

Restriction

- query set size control
- query set overlap control
- maximum order control
- partitioning
- cell suppression
- auditing

Modification

- data perturbation
- response perturbation
- data swapping (shuffling)
- random sample

Published Data Table

Name *	City	Age	Sex	Status	Post-traumatic stress disorder	Attempted suicide
White	Sydney	34	F	W	4.1	no
Scarlet	Dubbo	27	F	D	3.9	no
Brown	Sydney	45	M	M	4.3	no
Mustard	Perth	32	M	S	2.1	yes
Green	Ballina	76	M	M	4.8	no
Green	Darwin	32	F	M	4.6	no
Plum	Hobart	25	M	D	2.9	no
Mustard	Darwin	54	M	W	4.2	no
White	Dubbo	51	F	D	3.8	no
Peacock	Sydney	40	F	M	4.1	no
Black	Ballina	68	F	W	3.6	no
Violet	Dubbo	33	F	M	2.7	no
Aureate	Sydney	28	F	S	3.5	no

	ID - unique identifier ID={Name}
	QID - Quasi identifier QID={City,Age, Sex}
	Non-sensitive attributes
	Sensitive Attribute

* Strictly speaking, name itself can rarely be considered to be a unique identifier.

Attack Models

We can classify the main attack types into 2 broad categories:

1. **Linkage Attack Models:**
   1. **Record linkage,** where the intruder is able to link an individual to a record in the published data table.
   2. **Attribute linkage,** where the intruder is able to link an individual to a sensitive value in the published data table.
   3. **Table linkage,** where the intruder is able to link an individual to the published data table itself.

Attack Models

2. **Probabilistic attack**. The published table should provide the adversary with as little additional knowledge as possible, beyond what he/she already knew before seeing the table (background knowledge). *Probabilistic attack* occurs when the difference between the prior and the posterior knowledge is significant.

Table 2.1: Privacy models

Privacy Model	Attack Model			
	Record linkage	Attribute linkage	Table linkage	Probabilistic attack
k-Anonymity [201, 217]	✓			
MultiR k-Anonymity [178]	✓			
ℓ-Diversity [162]	✓	✓		
Confidence Bounding [237]		✓		
(α, k)-Anonymity [246]	✓	✓		
(X, Y)-Privacy [236]	✓	✓		
(k, e)-Anonymity [269]		✓		
(ϵ, m)-Anonymity [152]		✓		
Personalized Privacy [250]		✓		
t-Closeness [153]		✓		✓
δ-Presence [176]			✓	
(c, t)-Isolation [46]	✓			✓
ϵ-Differential Privacy [74]			✓	✓
(d, γ)-Privacy [193]			✓	✓
Distributional Privacy [33]			✓	✓

Record Linkage

The intruder is able to link an individual to a record in the published data table.

In published data tables, Unique Identifiers (UIs) are usually removed, so record linkage typically relies on QIDs.

Suppose that the individual A the adversary is after has a value *qid* of the QID, where the value qid is known to the adversary.

qid identifies a group of records in the table. If the size of the group is 1, we have record linkage.

If the size of the group is more than 1, the adversary may still be able to uniquely identify A with the help of additional knowledge.

Record Linkage Example 1

Name *	City	Age	Sex	Status	Post-traumatic stress disorder	Attempted suicide
White	Sydney	34	F	W	4.1	no
Scarlet	Dubbo	27	F	D	3.9	no
Brown	Sydney	45	M	M	4.3	no
Mustard	Perth	32	M	S	2.1	yes
Green	Ballina	76	M	M	4.8	no
Green	Darwin	32	F	M	4.6	no
Plum	Hobart	25	M	D	2.9	no
Mustard	Darwin	24	M	W	4.2	no
White	Dubbo	51	F	D	3.8	no
Peacock	Sydney	40	F	M	4.1	no
Black	Ballina	68	F	W	3.6	no
Violet	Dubbo	33	F	M	2.7	no
Aureate	Sydney	28	F	S	3.5	no

QID={City,Age,Sex}

ID(A) = Scarlet

qid=QID(A) = {Dubbo,27,F}

Record Linkage Example 2

Table 2.2: Original patient data

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

Table 2.3: External data

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

A hospital intends to release the Patient Data table (Table 2.2.) to a research centre.

The research centre already has access to an external data (Table 2.3.). Also, they know that every patient in Table 2.2. also has a record in Table 2.3.

What are ID, QID, non-sensitive and sensitive attributes in each table?

What can people from the research centre learn by linking these two tables?

K-anonymity

In her famous 2002 paper [5], Sweeny showed that 87% of respondents in 1990 US census (216,000,000) can be uniquely identified using only 3 attributes:

- ZIP code
- Date of Birth
- Gender

- In the same paper she famously demonstrated how linking different data sets can be used to compromise sensitive information about individuals.

K-anonymity

The National Association of Health Data Organizations (NAHDO) reported that 37 states in the USA have legislative mandates to collect hospital level data and that 17 states have started collecting ambulatory care data from hospitals, physicians offices, clinics, and so forth [2]. The leftmost circle in Figure 1 contains a subset of the fields of information, or *attributes*, that NAHDO recommends these states collect; these attributes include the patient's ZIP code, birth date, gender, and ethnicity.

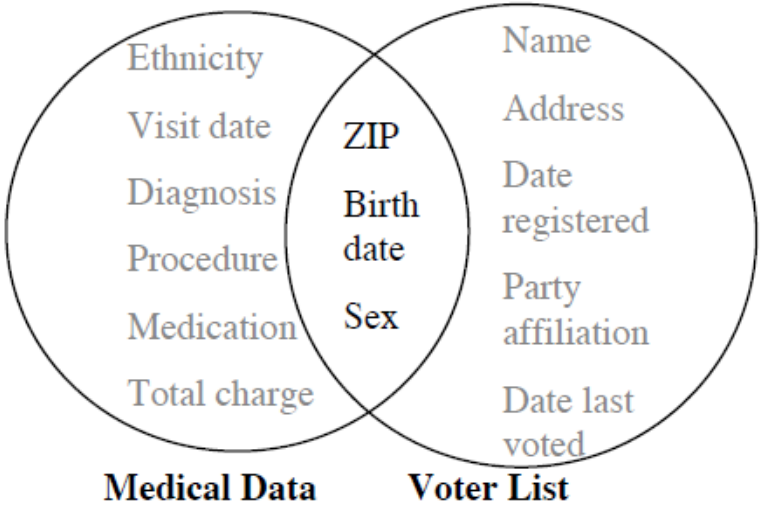

Figure 1 Linking to re-identify data

K-anonymity

In Massachusetts, the Group Insurance Commission (GIC) is responsible for purchasing health insurance for state employees. GIC collected patient-specific data with nearly one hundred attributes per encounter along the lines of the those shown in the leftmost circle of Figure 1 for approximately 135,000 state employees and their families. Because the data were believed to be anonymous, GIC gave a copy of the data to researchers and sold a copy to industry [3].

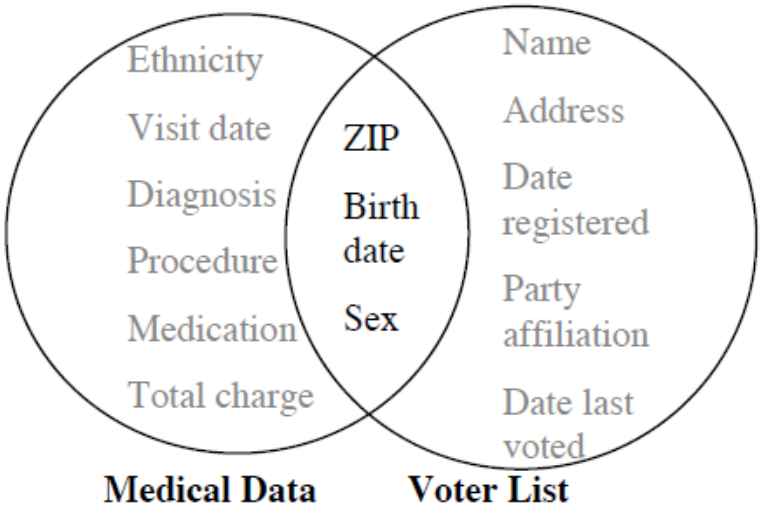

Figure 1 Linking to re-identify data

K-anonymity

For twenty dollars I purchased the voter registration list for Cambridge Massachusetts and received the information on two diskettes [4]. The rightmost circle in Figure 1 shows that these data included the name, address, ZIP code, birth date, and gender of each voter. This information can be linked using ZIP code, birth date and gender to the medical information, thereby linking diagnosis, procedures, and medications to particularly named individuals.

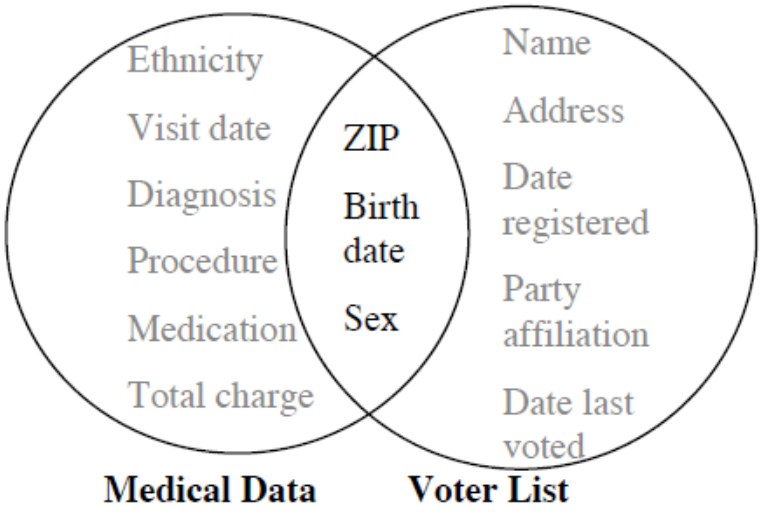

Figure 1 Linking to re-identify data

K-anonymity

For example, William Weld was governor of Massachusetts at that time and his medical records were in the GIC data. Governor Weld lived in Cambridge Massachusetts. According to the Cambridge Voter list, six people had his particular birth date; only three of them were men; and, he was the only one in his 5-digit ZIP code.

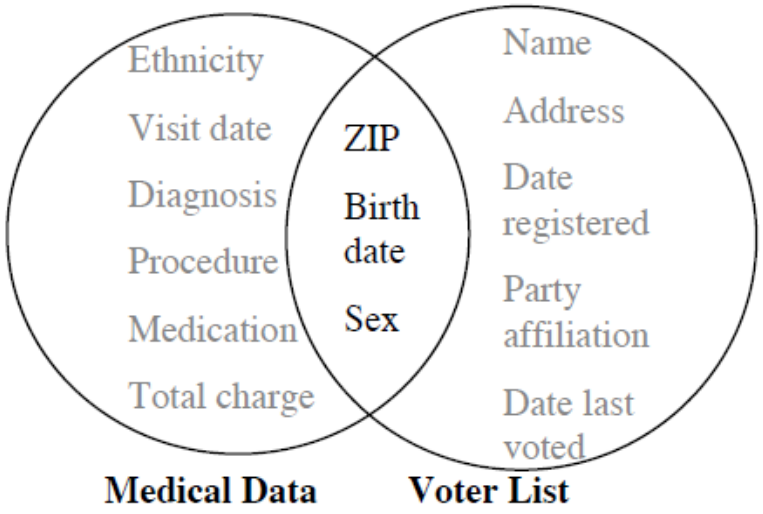

Figure 1 Linking to re-identify data

K-anonymity

Together with Samarati, [4,5,6], Sweeny proposed k-anonymity to prevent record linkage:

For each value *qid* of QID that exist in the data table, there are at least k record having value *qid* in QID.

A table satisfying this requirement is called k-anonymous. In such a table, a probability of successfully linking a record to another table on QID is at most $\frac{1}{k}$.

K-anonymity Example 1 [1]

Table 2.2: Original patient data

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

Table 2.3: External data

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

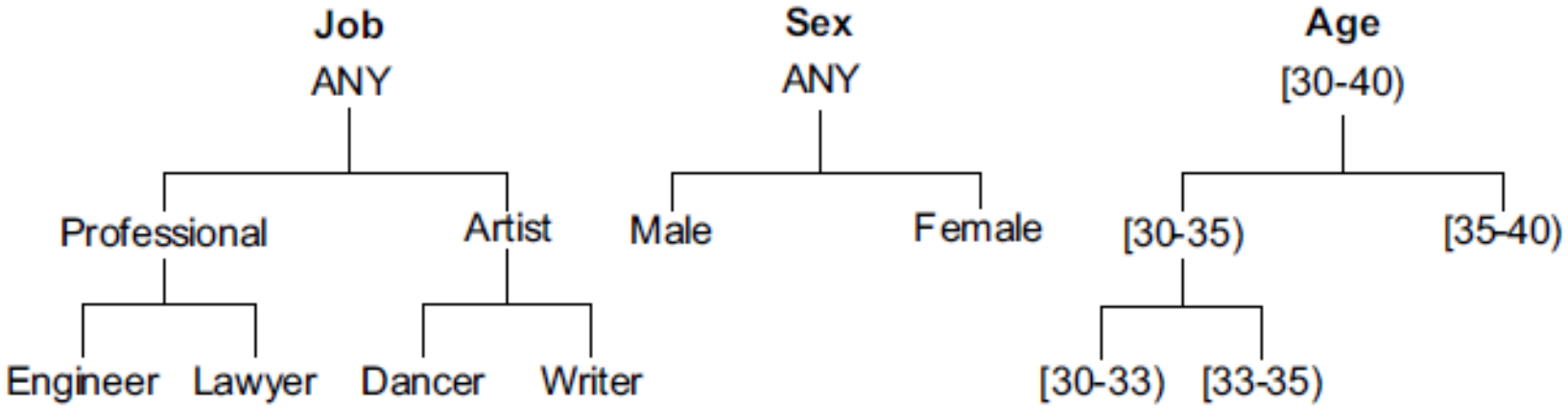

FIGURE 2.1: Taxonomy trees for *Job*, *Sex*, *Age*

K-anonymity Example 1 [1]

Table 2.2: Original patient data

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

Table 2.3: External data

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

Table 2.4: 3-anonymous patient data

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

Table 2.5: 4-anonymous external data

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)

K-anonymity and Quasi-identifiers

We assume that QID is known to the adversary.

How to select a QID?

- Include all attributes that could be known to the adversary into the QID. This would increase privacy but decrease utility of the data.
- Use multiple QIDs. This is useful if the data manager (owner, holder) knows the tables the adversary may use for linking.

K-anonymity and Multiple QIDs

Example 1 [1].

Data manager wants to publish a table T=(A,B,C,D,S), where S is a sensitive attribute. Data user already has access to two other tables, T_1=(A,B,X) and T_2=(C,D,Y). Then data manager can use two QUIs to provide k-anonymity: QID_1=(A,B) and QID_2=(C,D).

Question 1: Does k-anonymity on QID=(A,B,C,D) provide k-anonymity on QID_1=(A,B) and QID_2=(C,D)?

Question 2: Does k-anonymity on QID_1=(A,B) and QID_2=(C,D) provide k-anonymity on QID=(A,B,C,D)?

Question 3: Let QID' $\subseteq$ QID. Does k-anonymity on QID provide k-anonymity on QID'? What about the other way around?

K-anonymity and Multiple Records per Individual

So far we have assumed that each individual corresponds to at most one record in the data table.

However, it is possible that the table contains multiple records per an individual. Such tables are typically obtained by joining multiple tables.

Example [1].

Consider a table Patient=(PID, Age, Gender, Disease) and let QID={ Age, Gender}. Note that a patient can have more that one disease, therefore more than one record in the table. Thus a group of k records with a same qid may contain less than k patients.

Linkage Attack Models

1. **Linkage Attack Models:**
   1. **Record linkage,** where the intruder is able to link an individual to a record in the published data table.
   2. **Attribute linkage,** where the intruder is able to link an individual to a sensitive value in the published data table.
   3. **Table linkage,** where the intruder is able to link an individual to the published data table itself.

Table 2.1: Privacy models

Privacy Model	Attack Model			
	Record linkage	Attribute linkage	Table linkage	Probabilistic attack
k-Anonymity [201, 217]	✓			
MultiR k-Anonymity [178]	✓			
ℓ-Diversity [162]	✓	✓		
Confidence Bounding [237]		✓		
(α, k)-Anonymity [246]	✓	✓		
(X, Y)-Privacy [236]	✓	✓		
(k, e)-Anonymity [269]		✓		
(ϵ, m)-Anonymity [152]		✓		
Personalized Privacy [250]		✓		
t-Closeness [153]		✓		✓
δ-Presence [176]			✓	
(c, t)-Isolation [46]	✓			✓
ϵ-Differential Privacy [74]			✓	✓
(d, γ)-Privacy [193]			✓	✓
Distributional Privacy [33]			✓	✓

Attribute Linkage

Even if the adversary is not able to perform record linkage, they still may be able to perform attribute linkage and disclose the sensitive attribute value of an individual, or significant information about the sensitive value.

Attribute Linkage Example [1]

Table 2.4: 3-anonymous patient data

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

Table 2.5: 4-anonymous external data

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)

l-diversity

In order to prevent attribute linkage, Machanavajjhala et al. in 2007 proposed the l-diversity privacy model.

Informally, ***l-diversity*** requires that every qid equivalence group contains at least l well-represented values in each sensitive attribute.

Distinct l-diversity

The simplest version of l-diversity is ***distinct l-diversity***, where every qid equivalence group contains at least l distinct values in each sensitive attribute.

Note that distinct l-diversity satisfies k-anonymity for k = l.

Distinct l-diversity cannot prevent probabilistic attack as the frequency of different sensitive values can vary greatly.

Entropy l-diversity

Entropy l-diversity requires that for every qid equivalence group and each sensitive attribute we have

$$-\sum_{s \epsilon S} P(qid, s) \lg P(qid, s)) \geq \log(l)$$

where S is an actual domain of the sensitive attribute (the set of values that actually exist in the data table), and P(qid,s) is the fraction of records in qid equivalence group having sensitive value s.

Entropy revision

$$-\sum_{s \epsilon S} P(qid, s) \lg P(qid, s)) = \sum_{s \epsilon S} P(qid, s) \lg \frac{1}{P(qid, s))} \geq \log(l)$$

Entropy l-diversity Example [1]

$$-\sum_{s \epsilon S} P(qid, s) \lg P(qid, s)) \geq \log(l)$$

Example: Calculate l for Distinct and Entropy l-diversity for the following table.

Table 2.4: 3-anonymous patient data

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

Entropy l-diversity Example [1]

Solution:

$$-\sum_{s \epsilon S} P(qid, s) \lg P(qid, s)) \geq \log(l)$$

Table 2.4: 3-anonymous patient data

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

For the first equivalence class with qid={Professional, Male, [35-40]} we have: $-\frac{2}{3}\lg\frac{2}{3} - \frac{1}{3}\lg\frac{1}{3} \geq \lg(l)$. Therefore, $l \leq 1.9$

For the second equivalence class with
qid={Artist, Female, [30-35]}
we have: $-\frac{3}{4}\lg\frac{3}{4} - \frac{1}{4}\lg\frac{1}{4} \geq \lg(l)$. Therefore, $l \leq 1.8$.

Putting the two inequalities together we get $l \leq 1.8$.

Table 2.1: Privacy models

Privacy Model	Attack Model			
	Record linkage	Attribute linkage	Table linkage	Probabilistic attack
k-Anonymity [201, 217]	✓			
MultiR k-Anonymity [178]	✓			
ℓ-Diversity [162]	✓	✓		
Confidence Bounding [237]		✓		
(α, k)-Anonymity [246]	✓	✓		
(X, Y)-Privacy [236]	✓	✓		
(k, e)-Anonymity [269]		✓		
(ϵ, m)-Anonymity [152]		✓		
Personalized Privacy [250]		✓		
t-Closeness [153]		✓		✓
δ-Presence [176]			✓	
(c, t)-Isolation [46]	✓			✓
ϵ-Differential Privacy [74]			✓	✓
(d, γ)-Privacy [193]			✓	✓
Distributional Privacy [33]			✓	✓

ε-differential privacy

Dwork, 2006, proposes different privacy notion: the risk to the record owner's privacy should not substantially increase as a result of participating in a statistical database.

Instead of comparing the prior probability and the posterior probability before and after accessing the published data, Dwork proposes to compare the risk with and without the record owner's data in the published data.

ε-differential privacy

Consequently, the privacy model called ε-differential privacy ensures that the removal or addition of a single database record does not significantly affect the outcome of any analysis. It follows that no risk is incurred by joining different databases.

Based on the same intuition, if a record owner does not provide his/her actual information to the data holder, it will not make much difference in the result of the anonymization algorithm.

ε-differential privacy

The following is a formal definition of ε-differential privacy.

A randomized function F ensures ε-differential privacy if for all data sets T_1 and T_2 differing on at most one record,

$$\left| \ln \frac{P[F(T_1) = S]}{P[F(T_2) = S]} \right| \leq \varepsilon$$

For all S ∈ Range(F), where Range(F) is the set of possible outputs of the randomized function F.

ε-differential privacy

Although ε-differential privacy does not prevent record and attribute linkages studied in earlier chapters, it assures record owners that they may submit their personal information to the database securely in the knowledge that nothing, or almost nothing, can be discovered from the database with their information that could not have been discovered without their information.

ε-differential privacy

Dwork proves that if the number of queries is sub-linear in n, the noise to achieve differential privacy is bounded by o(√n),where n is the number of records in the database.

Dwork further shows that the notion of differential privacy is applicable to both interactive and non-interactive query models.

References

[1] B. C. M. Fung, K. Wang, A. W.-C. Fu and P. S. Yu, *Introduction to Privacy-Preserving Data Publishing - Concepts and Techniques*, CRC Press, Tylor & Francis Group, 2011.

[2] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM* Transactions on Knowledge Discovery from Data (TKDD), 1(1), March 2007.

[3] C. Dwork. Differential privacy. In Proc. of the 33rd International Colloquium on Automata, Languages and Programming (ICALP), pages 1–12, Venice, Italy, 2006.

References

[4] P. Samarati and L. Sweeney. Generalizing data to provide anonymity when disclosing information. In *Proc. of the 17th ACM SIGACTSIGMOD-SIGART Symposium on Principles of Database Systems (PODS)*, page 188, Seattle, WA, June 1998.

[5] L. Sweeney, "k-anonymity: a model for protecting privacy", *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, **10**(5), 557-570, 2002.

[6] P. Samarati. Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 13(6):1010–1027, 2001.