Minimality Attack in Privacy Preserving Data Publishing

Raymond Chi-Wing Wong (the Chinese University of Hong Kong)
Ada Wai-Chee Fu (the Chinese University of Hong Kong)
Ke Wang (Simon Fraser University)
Jian Pei (Simon Fraser University)

Prepared by Raymond Chi-Wing Wong
Presented by Raymond Chi-Wing Wong

Outline

1. Introduction
   - k-anonymity
   - l-diversity
2. Enhanced model
   - Weaknesses of l-diversity
   - m-confidentiality
3. Algorithm
4. Experiment
5. Conclusion

Minimize information loss, which gives rise to a new attack called **Minimality Attack**.

1. K-Anonymity

Patient	Gender	Address	Birthday	Cancer
Raymond	Male	Hong Kong	29 Jan	None
Peter	Male	Shanghai	16 July	Yes
Kitty	Female	Hong Kong	21 Oct	None
Mary	Female	Hong Kong	8 Feb	None

Release the data set to **public**

Gender	Address	Birthday	Cancer
Male	Hong Kong	29 Jan	None
Male	Shanghai	16 July	Yes
Female	Hong Kong	21 Oct	None
Female	Hong Kong	8 Feb	None

1. K-Anonymity

QID (quasi-identifier)

Patient	Gender	Address	Birthday	Cancer
Raymond	Male	Hong Kong	29 Jan	None
Peter	Male	Shanghai	16 July	Yes
Kitty	Female	Hong Kong	21 Oct	None
Mary	Female	Hong Kong	8 Feb	None

Knowledge 2

I also know Peter with (Male, Shanghai, 16 July)

Release the data set to **public**

Knowledge 1

Gender	Address	Birthday	Cancer
Male	Hong Kong	29 Jan	None
Male	Shanghai	16 July	Yes
Female	Hong Kong	21 Oct	None
Female	Hong Kong	8 Feb	None

Combining Knowledge 1 and Knowledge 2, we may deduce the ORIGINAL person.

1. K-Ano

2-anonymity: to generate a data set such that each possible QID value appears at least TWO times.

QID (quasi-identifier)

Patient	Gender	Address	Birthday	Cancer
Raymond	Male	Hong Kong	29 Jan	None
Peter	Male	Shanghai	16 July	Yes
Kitty	Female	Hong Kong	21 Oct	None
Mary	Female	Hong Kong	8 Feb	None

Knowledge 2

I also know Peter with (Male, **Asia**, 16 July)

Release the data set to **public**

In the released data set,
each possible QID value (Gender, Address, Birthday) appears at least TWO times.

Knowledge 1

	Address	Birthday	Cancer
Male	**Asia**	*	None
Male	**Asia**	*	Yes
Female	Hong Kong	*	None
Female	Hong Kong	*	None

Combining Knowledge 1 and Knowledge 2, we **CANNOT** deduce the ORIGINAL person.

This data set is 2-anonymous

1. K-anonymity

- We have discussed the traditional model of k-anonymity
- Does this model really preserve “privacy”?

Gender	Address	Birthday	Cancer
Male	**Asia**	*	Yes
Male	**Asia**	*	Yes
Female	Hong Kong	*	None
Female	Hong Kong	*	None

1. l-diversity

Patient	Gender	Address	Birthday	Cancer
Raymond	Male	Hong Kong	29 Jan	None
Peter	Male	Shanghai	16 July	Yes
Kitty	Female	Shanghai	21 Oct	None
Mary	Female	Hong Kong	8 Feb	None

Release the data set to **public**

Gender	Address	Birthday	Cancer
Male	Hong Kong	29 Jan	None
Male	Shanghai	16 July	Yes
Female	Shanghai	21 Oct	None
Female	Hong Kong	8 Feb	None

1. l-diversity

Patient	Gender	Address	Birthday	Cancer
Raymond	Male	Hong Kong	29 Jan	None
Peter	Male	Shanghai	16 July	Yes
Kitty	Female	Shanghai	21 Oct	None
Mary	Female	Hong Kong	8 Feb	None

Knowledge 2

I also know Peter with (Male, Shanghai, 16 July)

Release the data set to **public**

Knowledge 1

Gender	Address	Birthday	Cancer
Male	Hong Kong	29 Jan	None
Male	Shanghai	16 July	Yes
Female	Shanghai	21 Oct	None
Female	Hong Kong	8 Feb	None

Combining Knowledge 1 and Knowledge 2, we may deduce the disease of Peter.

1. l-diversity

Patient	Gender	Address	Birthday	Cancer
Raymond	Male	Hong Kong	29 Jan	None
Peter	Male	Shanghai	16 July	Yes
Kitty	Female	Shanghai	21 Oct	None
Mary	Female	Hong Kong	8 Feb	None

Knowledge 2

I also know Peter with (Male, Shanghai, 16 July)

Release the data set to **public**

Knowledge 1

Gender	Address	Birthday	Cancer
Male	Hong Kong	29 Jan	None
Male	Shanghai	16 July	Yes
Female	Shanghai	21 Oct	None
Female	Hong Kong	8 Feb	None

1. l-diversity

Simplified 2-diversity: to generate a data set such that each individual is linked to “cancer” with probability at most 1/2

Patient	Gender	Address	Birthday	Cancer
Raymond	Male	Hong Kong	29 Jan	None
Peter	Male	Shanghai	16 July	Yes
Kitty	Female	Shanghai	21 Oct	None
Mary	Female	Hong Kong	8 Feb	None

Knowledge 2

I also know Peter with (Male, Shanghai, 16 July)

Now, we cannot deduce “Peter” suffered from “Cancer”

Release the data set to **public**

These two tuples form an **equivalence class**.

Knowledge 1

	Address	Birthday	Cancer
	Hong Kong	*	None
*	Shanghai	*	Yes
*	Shanghai	*	None
*	Hong Kong	*	None

Combining Knowledge 1 and Knowledge 2, we **CANNOT** deduce the disease of Peter.

This data set is 2-diverse

2.1 Weakness of l-diversity

- We have discussed l-diversity
- Does this model really preserve "privacy"?
- No.

2.1 Weakness of l-diversity

Simplified 2-diversity: to generate a data set such that each individual is linked to “cancer” with probability at most 1/2

Patient	QID	Cancer
Raymond	q1	None
Peter	q2	Yes
Kitty	q3	None
Mary	q4	None

Knowledge 2

I also know Peter with (Male, Shanghai, 16 July)

Release the data set to **public**

Knowledge 1

QID	Cancer
Q1	None
Q2	Yes
Q2	None
Q1	None

2.1 Weakness of l-diversity

Simplified 2-diversity: to generate a data set such that each individual is linked to “cancer” with probability at most 1/2

QID	Cancer
q1	None
q2	Yes
q3	None
q4	None

Release the data set to **public**

QID	Cancer
Q1	None
Q2	Yes
Q2	None
Q1	None

l-diversity

Simplified 2-diversity: to generate a data set such that each individual is linked to "cancer" with probability at most 1/2

e.g.1

QID	Cancer
q1	Yes
q1	None
q2	Yes
q2	None
q2	None
q2	None

e.g.2

QID	Cancer
q1	Yes
q1	Yes
q2	None
q2	None
q2	None
q2	None

Does NOT satisfy 2-diversity

Satisfies 2-diversity

Release the data set to **public**

QID	Cancer
q1	Yes
q1	None
q2	Yes
q2	None
q2	None
q2	None

QID	Cancer
Q	Yes
Q	Yes
Q	None
Q	None
q2	None
q2	None

Satisfies 2-diversity

Satisfies 2-diversity

l-diversity

Simplified 2-diversity: to generate a data set such that each individual is linked to “cancer” with probability at most 1/2

e.g.1

QID	Cancer
q1	Yes
q1	None
q2	Yes
q2	None
q2	None
q2	None

e.g.2

QID	Cancer
q1	Yes
q1	Yes
q2	None
q2	None
q2	None
q2	None

Does NOT satisfy 2-diversity

Satisfies 2-diversity

Same set of sensitive values (i.e. Cancer)

Same set of QID values

Release the data set to **public**

QID	Cancer
q1	Yes
q1	None
q2	Yes
q2	None
q2	None
q2	None

QID	Cancer
Q	Yes
Q	Yes
Q	None
Q	None
q2	None
q2	None

Different released data sets!

Why?

The anonymization algorithm tries to **minimize** the generalization steps.

Satisfies 2-diversity

Satisfies 2-diversity

l-diversity

Simplified 2-diversity: to generate a data set such that each individual is linked to “cancer” with probability at most 1/2

e.g.1

e.g.2

QID	Cancer
q1	Yes
q1	Yes
q2	None
q2	None
q2	None
q2	None

Release the data set to **public**

QID	Cancer
Q	Yes
Q	Yes
Q	None
Q	None
q2	None
q2	None

2.1 W l-diversity

Simplified 2-diversity: to generate a data set such that each individual is linked to “cancer” with probability at most 1/2

QID	Cancer
q1	Yes
q1	Yes
q2	None
q2	None
q2	None
q2	None

QID	Cancer
Q	Yes
Q	Yes
Q	None
Q	None
q2	None
q2	None

Simplified 2-diversity: to generate a data set such that each individual is linked to “cancer” with

e… of l-diversity

QID	Cancer
q1	Yes
q1	Yes
q2	None
q2	None
q2	None
q2	None

Knowledge 2

I also know Peter with QID = (q1)

Knowledge 3

I also know that there are two q1 values and four q2 values in the table.

Knowledge 4

The anonymization algorithm tries to **minimize** the generalization steps for 2-diversity

I will think in the following way.

Knowledge 1

QID	Cancer
Q	Yes
Q	Yes
Q	None
Q	None
q2	None
q2	None

Poss. 1

QID	Cancer
q1	Yes
q1	Yes
q2	None
q2	None
q2	None
q2	None

Poss. 2

QID	Cancer
q2	Yes
q2	Yes
q1	None
q1	None
q2	None
q2	None

Poss. 3

QID	Cancer
q1	Yes
q2	Yes
q1	None
q2	None
q2	None
q2	None

of l-diversity

Simplified 2-diversity: to generate a data set such that each individual is linked to "cancer" with

Knowledge 2

I also know Peter with QID = (q1)

Knowledge 3

I also know that there are two q1 values and four q2 values in the table.

Knowledge 4

The anonymization algorithm tries to **minimize** the generalization steps for 2-diversity

I will think in the following way.

Suppose the original table is Poss. 2.

- TWO q1 values are NOT linked to "Yes".
- FOUR q2 values are linked to TWO "Yes"'s.

The original table satisfies 2-diversity.

There is NO need to generalize q1 and q2 to Q.

Knowledge 1

QID	Cancer
Q	Yes
Q	Yes
Q	None
Q	None
q2	None
q2	None

Poss. 1

QID	Cancer
q1	Yes
q1	Yes
q2	None
q2	None
q2	None
q2	None

Poss. 2

QID	Cancer
q2	Yes
q2	Yes
q1	None
q1	None
q2	None
q2	None

Poss. 3

QID	Cancer
q1	Yes
q2	Yes
q1	None
q2	None
q2	None
q2	None

Simplified 2-diversity: to generate a data set such that each individual is linked to “cancer” with

e… of l-diversity

Suppose the original table is Poss. 3.

- TWO q1 values are linked to ONE “Yes”.
- FOUR q2 values are linked to ONE “Yes”.

The original table satisfies 2-diversity.

There is NO need to generalize q1 and q2 to Q.

Knowledge 2

I also know Peter with QID = (q1)

Knowledge 3

I also know that there are two q1 values and four q2 values in the table.

…wledge 4

…mization algorithm tries to **minimize** the ge…ion steps for 2-diversity

I will think in the follo…ng way.

Knowledge 1

QID	Cancer
Q	Yes
Q	Yes
Q	None
Q	None
q2	None
q2	None

Poss. 1

QID	Cancer
q1	Yes
q1	Yes
q2	None
q2	None
q2	None
q2	None

Poss. 2

QID	Cancer
q2	Yes
q2	Yes
q1	None
q1	None
q2	None
q2	None

Poss. 3

QID	Cancer
q1	Yes
q2	Yes
q1	None
q2	None
q2	None
q2	None

e … of l-diversity

I deduce that the original table MUST be Poss. 1.

This person o MUST suffer From Cancer.

That is, P(o is linked to Cancer | Knowledge) = 1

This attack is called **Minimality Attack**.

Knowledge 2

I also know Peter with QID = (q1)

Knowledge 3

I also know that there are two q1 values and four q2 values in the table.

Knowledge 4

The anonymization algorithm tries to **minimize** the generalization steps for 2-diversity

I will think in the following way.

Knowledge 1

QID	Cancer
Q	Yes
Q	Yes
Q	None
Q	None
q2	None
q2	None

Poss. 1

QID	Cancer
q1	Yes
q1	Yes
q2	
q2	
q2	
q2	

Poss. 2

Poss. 3

m-confidentiality (where m = l)

Problem: to generate a data set which satisfies the following.

for each individual o,

P(o is linked to Cancer | Knowledge) <= 1/l

2.2 Minimality Attack

- Suppose A is the anonymization algorithm which tries to **minimize** the generalization steps for l-diversity.
  We call this the **minimality principle**.
- Let table T* be a table generated by A and T* satisfies l-diversity.
- Then, for any equivalence class E in T*,
  - there is no specialization (reverse of generalization) of the QID's in E which results in another table T' which also satisfies l-diversity.

2.2 Mi ttack

QID	Cancer
q1	Yes
q1	Yes
q2	None
q2	None
q2	None
q2	None

Does NOT satisfy 2-diversity

QID	Cancer
Q	Yes
Q	Yes
Q	None
Q	None
q2	None
q2	None

Satisfies 2-diversity

m-confidentiality (where m = l)

Problem: to generate a data set which satisfies the following.

for each individual o,
P(o is linked to Cancer | Knowledge) <= 1/l

2.3 Gener

- General Case
  - One special case was illustrated where

    P(o is linked to Cancer | Knowledge) = 1

  - In general, the computation of

    P(o is linked to Cancer | Knowledge)

    needs more sophisticated analysis.

2.3 General Formula (global recoding)

- P(o is linked to Cancer | Knowledge)
  - Try all possible cases
  - Consider a case
    - Consider o is in an equivalence class E
    - Suppose there are j tuples in E linked to Cancer
    - Proportion of tuples with Cancer = j/|E|
- P(o is linked to Cancer | Knowledge)
  $= \sum_{j=1}^{|E|} P(\text{no. of sensitive tuples} = j \mid \text{Knowledge}) \times j/|E|$

The derivation is accompanied by some exclusion of some possibilities by the adversary because of the minimality notion.

2.3 An Enhanced Model

- NP-hardness
  - Transform an NP-complete problem
    to this enhanced model (m-confidentiality)

  - NP-complete Problem:

    Exact Cover by 3-Sets(X3C)
    Given a set X with $|X| = 3q$ and a collection C of 3-element subsets of X. Does C contain an exact cover for X, i.e. a subcollection $C' \subseteq C$ such that every element of X occurs in exactly one member of C'?

2.4 General Model

- In addition to l-diversity, all existing models do not consider **Minimality Attack**
- The tables generated by the existing algorithm which follows **minimality principle** and satisfies one of the following privacy requirements have a privacy breach.
- Existing Requirements
  - (c, l)-diversity
  - (α, k)-anonymity
  - t-closeness
  - (k, e)-anonymity
  - (c, k)-safety
  - Personalized Privacy
  - Sequential Releases

3. Algorithm

- **Minimality Attack** exists
  when the anonymization method considers the "minimization" of the generalization steps for l-diversity
- **Key Idea** of Our proposed algorithm:
  we do not involve any "minimization" of generalization steps for l-diversity in our proposed algorithm
- With this idea, minimality attack is NOT possible.

3. Algorithm

- Some previous works pointed out that
  - k-anonymity has a privacy breach
- However, k-anonymity has been successful in some practical applications
- When a data set is k-anonymized,
  - the chance of a large proportion of a sensitive tuple in any equivalence class is very likely reduced to a safe level
- Since k-anonymity does not reply on the sensitive attribute,
  - we make use of k-anonymity in our proposed algorithm and perform some *precaution* steps to prevent the attack by minimality

3. Algorithm

- **Step 1: k-anonymization**
  - From the given table T, generate a k-anonymous table T^k (where k is a user parameter)
- **Step 2: Equivalence Class Classification**
  - From T^k, determine two sets:
    - set V containing a set of equivalence classes which violate l-diversity
    - set L containing a set of equivalence classes which satisfy l-diversity
- **Step 3: Distribution Estimation**
  - For each E in L,
    find the proportion p_i of tuples containing the sensitive value
  - Generate a distribution D according to p_i values of all E's in L
- **Step 4: Sensitive Attribute Distortion**
  - For each E in V,
    - randomly pick a value p_E from distribution D
    - distort the sensitive value in E such that the proportion of sensitive values in E is equal to p_E

3. Algorithm

- **Theorem:** Our proposed algorithm generates m-confidential data set.

for each individual o,

P(o is linked to Cancer | Knowledge) <= 1/m

4. Experiments

- Real Data Set (Adults)
  - 9 attributes
  - 45,222 instances
  - Default:
    - l = 2
    - QID size = 8
  - m = l

4. Experiments

- Real example
- QID attributes: age, workclass, marital status
- Sensitive attribuute: education

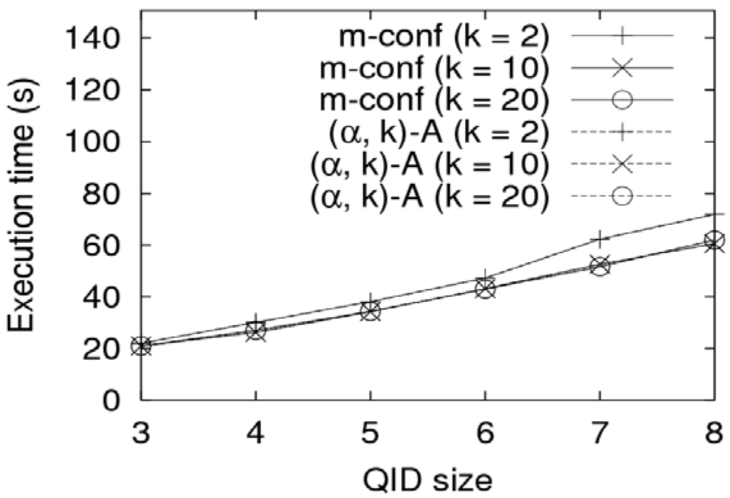
Age	Workclass	Marital Status	Education
80	Self-emp-not-inc	Married-spouse-absent	7th-8th
80	Private	Married-spouse-absent	HS-grad
80	private	Married-spouse-absent	HS-grad

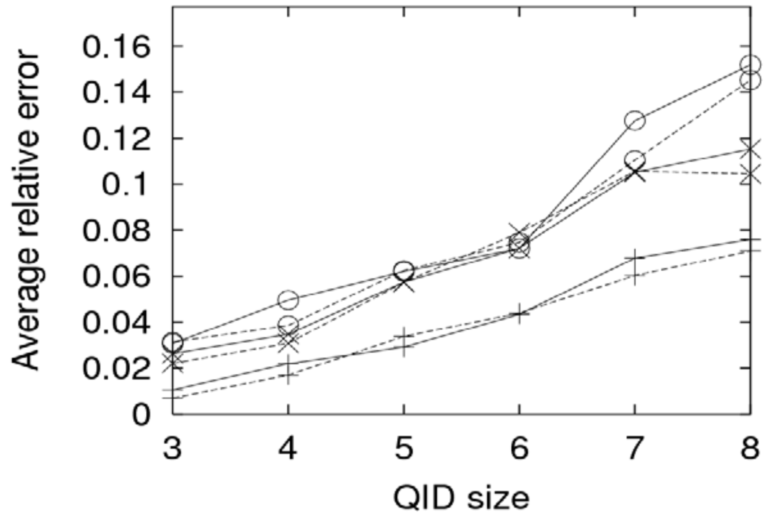
Age	Workclass	Marital Status	Education
80	With-pay	Married-spouse-absent	7th-8th
80	With-pay	Married-spouse-absent	HS-grad
80	private	Married-spouse-absent	HS-grad

4. Experiments

- Variation of QID size
- Compare our proposed algorithm with the algorithm which does not consider the minimality attack
- Measurement
  - Execution Time
  - Distortion after Anonymization

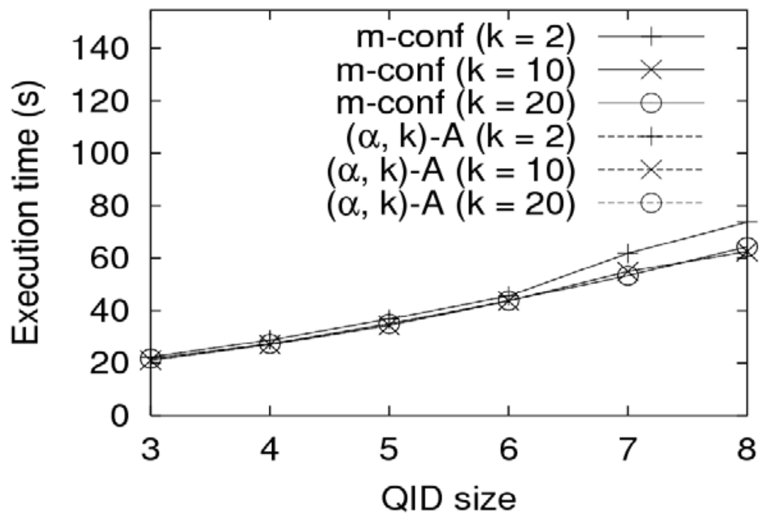
4. Experiments

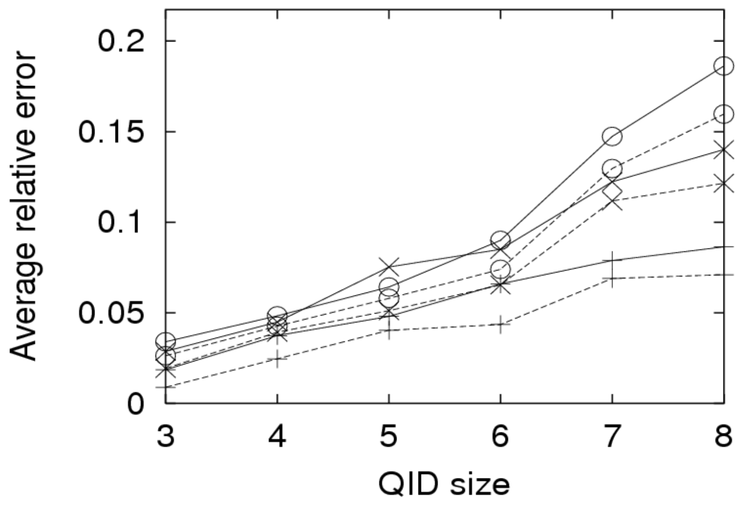




m = 2

4. Experiments





m = 10

5. Conclusion

- Minimality Attack
  - Exists in existing privacy models
- Derive Formulae of Calculating the Probability of privacy breaching
- Proposed algorithm
- Experiments

FAQ

Problem of 2-anonymity: to generate a data set such that each possible value appear at least two times

kness of l-diversity

QID	Cancer
q1	Yes
q2	Yes
q3	Yes
q3	None
q4	None
q4	None

QID	Cancer
Q	Yes
Q	Yes
q3	Yes
q3	None
q4	None
q4	None

Each possible value appears at least two times.

Problem: to find a data set which satisfies

1. ~~k-anonymity~~
2. α-deassociation requirement

Bucketization

QID	Cancer
q1	Yes
q2	Yes
q3	None
q4	None

Release the data set to **public**

QID	Cancer
Q1	Yes
Q2	Yes
Q2	None
Q1	None

QID	BID
q1	1
q4	1
q2	2
q3	2

BID	Cancer
1	Yes
1	None
2	Yes
2	None

di

QID	Disease
q1	Diabetics
q1	HIV
q1	Lung Cancer
q2	HIV
q2	Ulcer
q2	Alzhema
q2	Gallstones

QID	Disease
q1	Diabetics
q1	HIV
q1	HIV
q2	Lung Cancer
q2	Ulcer
q2	Alzhema
q2	Gallstones

(3, 3)-diversity

QID	Disease
q1	Diabetics
q1	HIV
q1	Lung Cancer
q2	HIV
q2	Ulcer
q2	Alzhema
q2	Gallstones

QID	Disease
Q	Diabetics
Q	HIV
Q	Lung Cancer
Q	HIV
q2	Ulcer
q2	Alzhema
q2	Gallstones

QID	Disease
q1	HIV
q1	none
q2	none
q2	none
q2	HIV
q2	HIV

QID	Disease
q1	HIV
q1	HIV
q2	none
q2	none
q2	none
q2	HIV

0.2-closeness

QID	Disease
q1	HIV
q1	none
q2	none
q2	none
q2	HIV
q2	HIV

QID	Disease
Q	HIV
Q	HIV
Q	none
q2	none
q2	none
q2	HIV

(k, e)-anonymity (k = 2)

(2, 5k)-anonymity

QID	Income
q1	30k
q1	20k
q2	30k
q2	20k
q2	40k

QID	Income
q1	30k
q1	20k
q2	30k
q2	20k
q2	40k

QID	Income
q1	30k
q1	30k
q2	20k
q2	10k
q2	40k

QID	Income
Q	30k
Q	30k
Q	20k
q2	10k
q2	40k

2)-

QID	Disease
q1	HIV
q1	none
q1	none
q2	HIV
q2	none
q2	none
q2	none
q2	none
q2	none
q2	none

QID	Disease
q1	HIV
q1	none
q1	none
q2	HIV
q2	none
q2	none
q2	none
q2	none
q2	none
q2	none
q2	none
q2	none

QID	Disease
q1	HIV
q1	HIV
q1	none
q2	none
q2	none
q2	none
q2	none
q2	none
q2	none
q2	none

QID	Disease
Q	HIV
Q	HIV
Q	none
Q	none
Q	none
Q	none
Q	none
Q	none
Q	none
q2	none
q2	none
q2	none

(0.6, 2)-safety

If an individual with q1 suffers from HIV, then another individual with q2 will suffer from HIV.

If an individual with q2 suffers from HIV, then another individual with q1 will suffer from HIV.

QID	Education	Guarding Node
q1	undergrad	none
q2	1st-4th	elementary
q2	undergrad	none

QID	Education	Guarding Node
q1	1st-4th	elementary
q2	undergrad	none
q2	undergrad	none

QID	Education
q1	undergrad
q2	1st-4th
q2	undergrad

QID	Education
Q	1st-4th
Q	undergrad
q2	undergrad

2-diversity for
Personalized privacy

Step 1

k-anonymization: From the given table T, generate a k-anonymous table T^k (where k is a user parameter)

Suppose k = 2

akness of l-diversity

QID	Cancer
Q	Yes
Q	Yes
q3	Yes
q3	None
q4	None
q4	None

Each possible value appears at least two times.

Step 2

Equivalence Class Classification: From T^k, determine two sets:

- set V containing a set of equivalence classes which violate 2-diversity
- set L containing a set of equivalence classes which satisfy 2-diversity

akness of l-diversity

V = { Q }

L = { q3 , q4 }

QID	Cancer
Q	Yes
Q	Yes
q3	Yes
q3	None
q4	None
q4	None

This equivalence class contains more than half sensitive tuples

This equivalence class contains at most half sensitive tuples

This equivalence class contains at most half sensitive tuples

Step 3

Distribution Estimation

- For each E in L,
  find the proportion p_i of tuples containing the sensitive value
- Generate a distribution D according to p_i values of all E's in L

V = { Q }

L = { q3 , q4 }

QID	Cancer
Q	Yes
Q	Yes
q3	Yes
q3	None
q4	None
q4	None

$p_i = 0.5$

$p_i = 0$

D = {0, 0.5}

In other words,

$Prob(p_i = 0) = 0.5$
$Prob(p_i = 0.5) = 0.5$

Step 4

Sensitive Attribute Distortion: For each E in V,

- randomly pick a value p_E from distribution D
- distort the sensitive value in E such that the proportion of sensitive values in E is equal to p_E

...l-diversity

V = { Q }

L = { q3 , q4 }

QID	Cancer
Q	Yes
Q	None
q3	Yes
q3	None
q4	None
q4	None

Distort the sensitive value such that p_E is equal to 0.5

Suppose p_E is equal to 0.5

$p_i = 0.5$

$p_i = 0$

D = {0, 0.5}

In other words,

$Prob(p_i = 0) = 0.5$

$Prob(p_i = 0.5) = 0.5$

Future Work

- An Enhanced Model of K-Anonymity
  - Try to find other possible enhanced models of K-Anonymity
- Minimality Attack in Privacy Preserving Data Publishing
  - Try to find other possible privacy breach which is based on the anonymization method

B.3 Algorithm

- **Step 1**: anonymize table T and generate a table T^k which satisfies k-anonymity
- **Step 2**:
  - find a set V of equivalence classes in T_k which violates α–deassociation
  - find a set L of equivalence classes in which satisfies α–deassociation
- **Step 3**:
  - generate distribution D on the proportion of sensitive value s of equivalence classes in L
- **Step 4**:
  - For each equivalence class E in V,
    - Randomly generate a number p_E from D
    - Distort the sensitive attribute of E such that the proportion of sensitive attribute is equal to p_E

B.1.2 K-

Problem: to generate a data set such that each possible value appears at least TWO times.

Customer	Gender	District	Birthday	Cancer
Raymond	Male	Shatin	29 Jan	None
Peter	Male	Fanling	16 July	Yes
Kitty	Female	Shatin	21 Oct	None
[illegible]	Female	Shatin	8 Feb	None

Two Kinds of Generalisations
1. Shatin→NT
2. 16 July→*

Release the data set to **public**

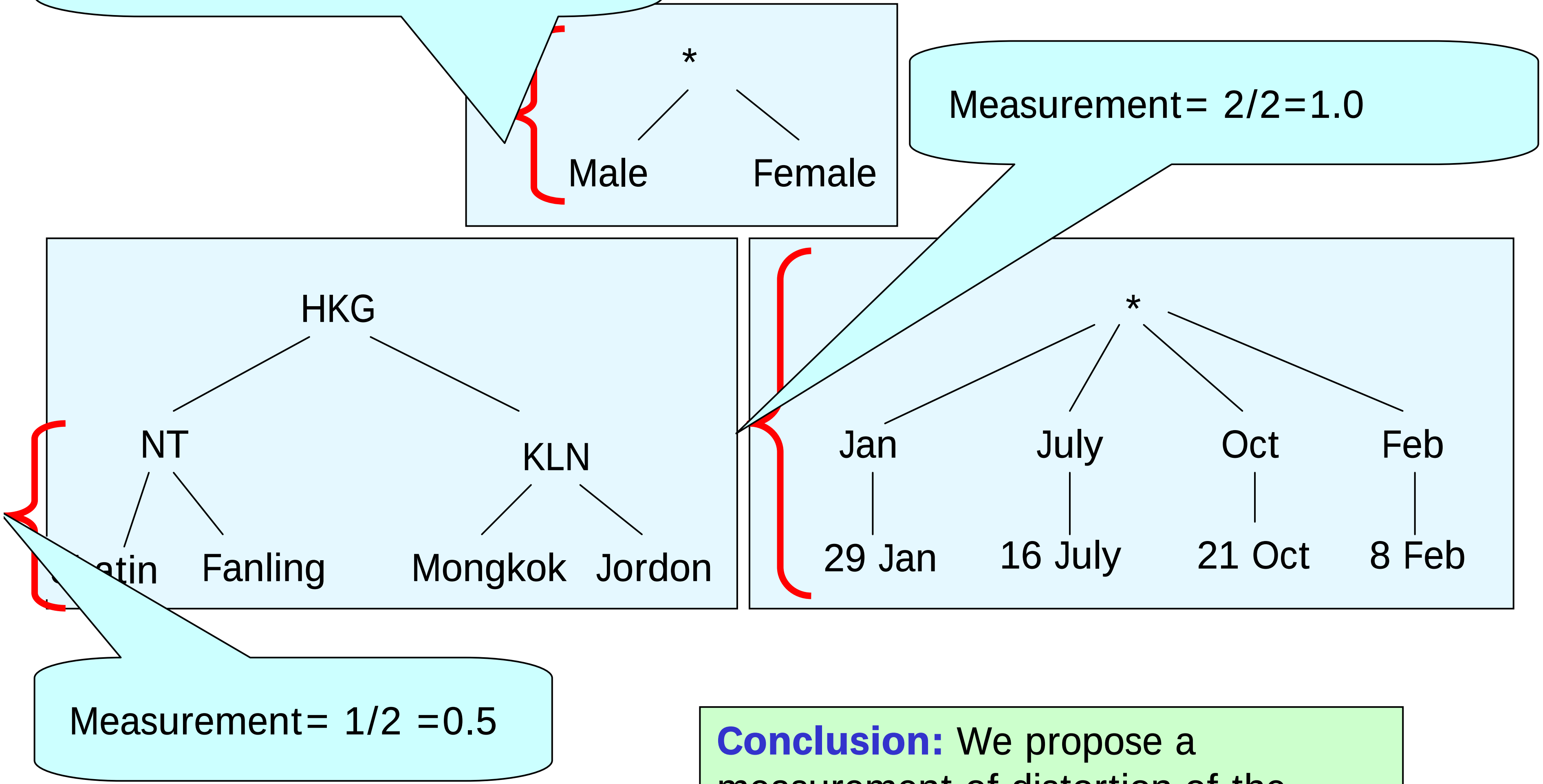
Gender	District	Birthday	Cancer
Male	**NT**	*	None
Male	**NT**	*	Yes
Female	Shatin	*	None
Female	Shatin	*	None

"Shatin→NT" causes LESS distortion than "16 July→*"

Question: how can we measure the distortion?

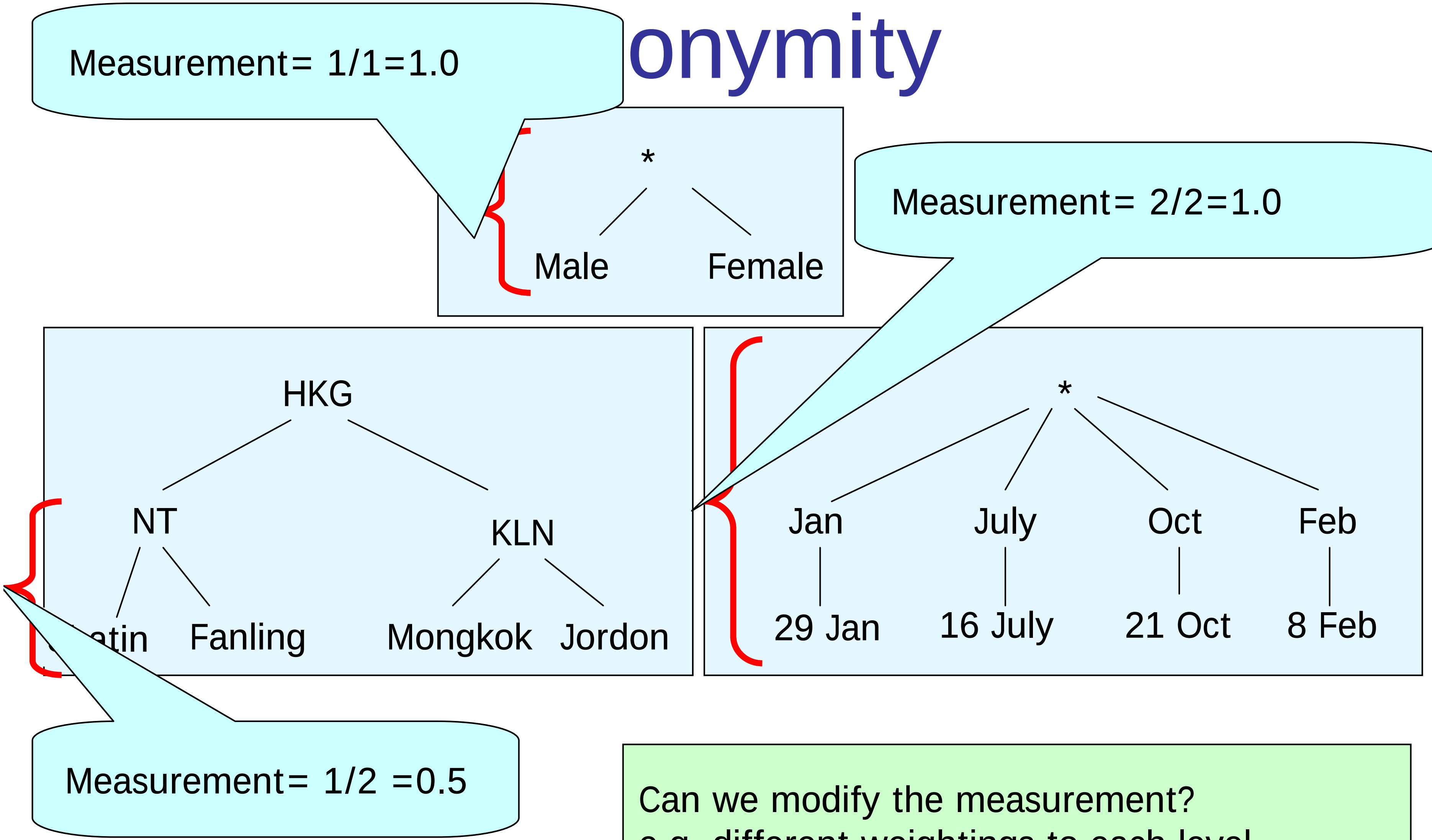
This data set is 2-anonymous

onymity



Conclusion: We propose a measurement of distortion of the modified/anonymized data.

onymity



Can we modify the measurement?
e.g. different weightings to each level

B.1.3 An Enhanced Model of K-Anonymity (Future Work)

Customer	Gender	District	Birthday	Cancer
Raymond	Male	Shatin	29 Jan	Yes
Peter	Male	Fanling	16 July	Yes
Kitty	Fe		1 Oct	None
Mary	Fe		Feb	None

Numerical Attribute?
Change Value?

Knowledge 2

I also know that there is a person with (Male, **NT**, 16 July)

Release the data set to **public**

Knowledge 1

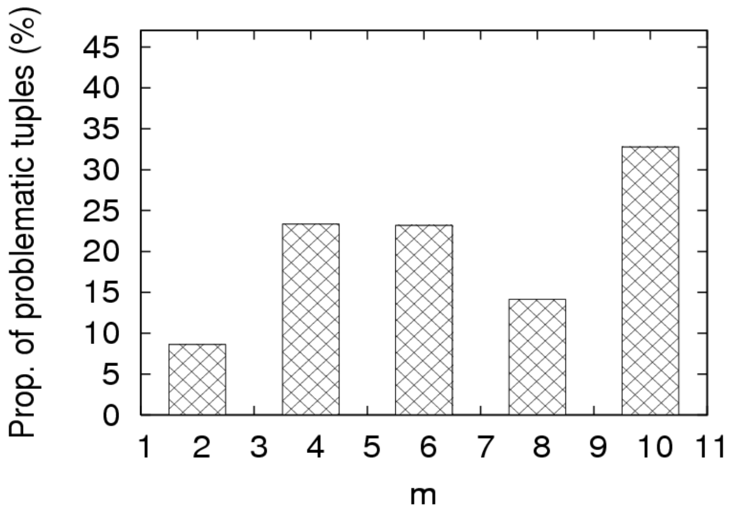
Gender	District	Birthday	Cancer
*	Shatin	*	Yes
*	**NT**	*	Yes
*	**NT**	*	None
*	Shatin	*	None

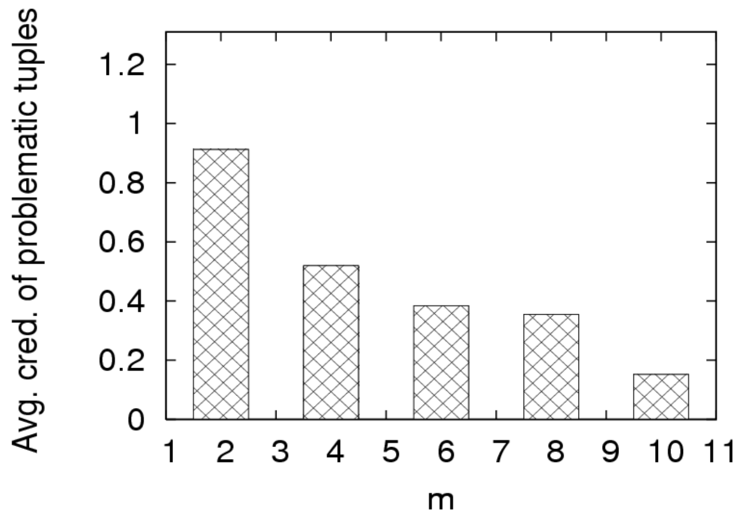
For each **equivalence class**, there are at most half records associated with "Cancer"

This is a user parameter.
In our problem, it is denoted by α (i.e. alpha)

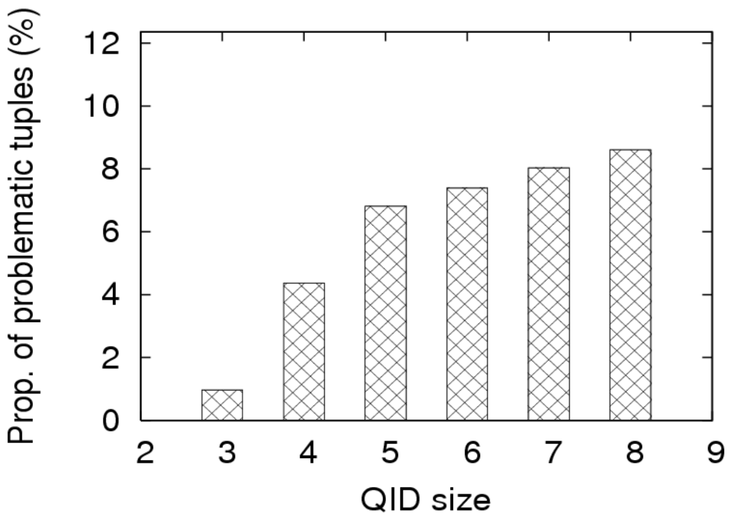
This data set is 2 anonymous

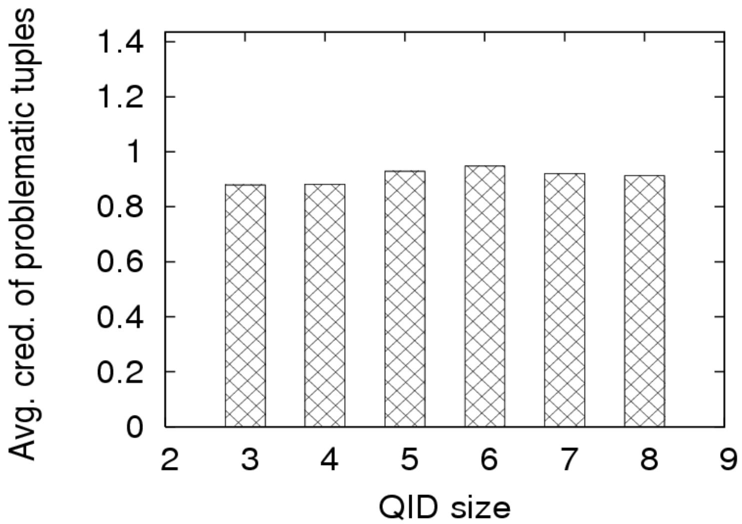
Experiments





Experiments





A.4 Experiments