k-Anonymity

Recuperação de Informação

Doutoramento em Engenharia Informática e Computadores

José Portêlo – 52096

Instituto Superior Técnico
Universidade Técnica de Lisboa

Outline

Introduction (1)

In the past, data was released in the form of tables and statistics (macrodata). Nowadays, there is an increasing demand for the release of specific data (microdata).

Introduction (1)

In the past, data was released in the form of tables and statistics (macrodata). Nowadays, there is an increasing demand for the release of specific data (microdata).

A naïve way to protect the anonymity of the entities is to remove or encrypt explicit identifiers. This, however, does not guarantee anonymity, as the remaining information can be linked using publicly available data.

Introduction (2)

A wide variety of private records describing each person's finances, interests and demographics is increasing every day, and they are being publicly distributed or sold.

Introduction (2)

A wide variety of private records describing each person's finances, interests and demographics is increasing every day, and they are being publicly distributed or sold.

This situation has raised concerns in different fields, as the amount of microdata that has been publicly released can be abused, thus compromising the privacy of individuals.

Motivation (1)

SSN	Name	Race	Date of birth	Sex	ZIP code	Marital status	Disease
		asian	64/04/12	F	94142	divorced	hypertension
		asian	64/09/13	F	94141	divorced	obesity
		asian	64/04/15	F	94139	married	chest pain
		asian	63/03/13	M	94139	married	obesity
		asian	63/03/18	M	94139	married	short breath
		black	64/09/27	F	94138	single	short breath
		black	64/09/27	F	94139	single	obesity
		white	64/09/27	F	94139	single	chest pain
		white	64/09/27	F	94141	widow	short breath

Name	Address	City	ZIP code	Date of birth	Sex	Marital status
.....	.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....	.....
Sue J. Doe	900 Market Street	San Francisco	*94142*	*64/02/12*	*F*	*divorced*
.....	.....	.....	.....	.....	.....	.....

Motivation (2)

Combining the information in both tables, it is possible to identify the medical condition of a unique individual.

Commonly used approaches to prevent linking attacks: sampling, swapping values, adding noise.

Many uses for microdata require it to be truthful, which means that none of these approaches can be applied.

k-Anonymity

It has been proposed as a way to protect the privacy of individuals while maintaining the information truthful.

k-Anonymity requirement

A quasi-identifier (QI) is the set of attributes included in the private table that are also publicly available.

Definition 1: k-anonymity requirement
Each release of data must be such that every combination of values of quasi-identifiers can be indistinctly matched to at least k individuals.

Since it is highly impractical and limiting to make assumptions on the datasets available, k-anonymity takes a safer approach.

k-Anonymity

Definition 2: k-anonymity
Let $T(A_1, ..., A_m)$ be a table, and QI be a quasi-identifier associated with it. T is said to satisfy k-anonymity with respect to QI iff each sequence of values in $T[QI]$ appears at least with k occurrences in $T[QI]$.

k-anonymity requires each quasi-identifier value to have at least k occurrences.

k-Anonymity

The *k*-anonymity proposal focuses on two techniques:

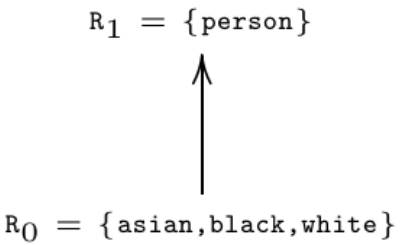
- Generalization

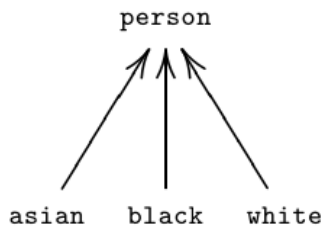
  Substituting the values of a given attribute with more general values.

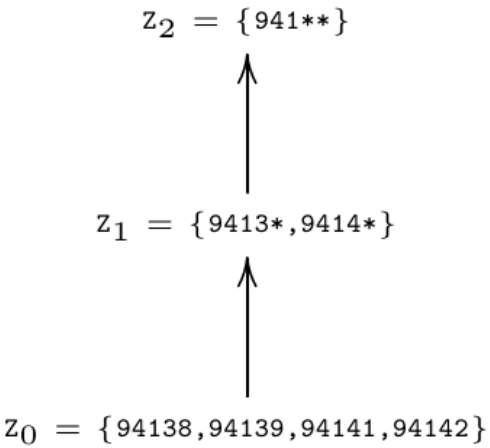
- Suppression

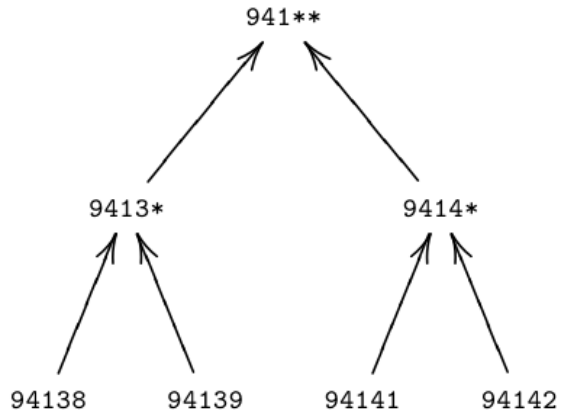
  Remove specific tuples (outliers) in order to reduce the amount of generalization required.

Generalization (1)









Generalization (2)

Race:R_0	ZIP code:Z_0
asian	94142
asian	94141
asian	94139
asian	94139
asian	94139
black	94138
black	94139
white	94139
white	94141

Race:R_1	ZIP code:Z_0
person	94142
person	94141
person	94139
person	94139
person	94139
person	94138
person	94139
person	94139
person	94141

Race:R_0	ZIP code:Z_1
asian	9414*
asian	9414*
asian	9413*
asian	9413*
asian	9413*
black	9413*
black	9413*
white	9413*
white	9414*

Race:R_1	ZIP code:Z_1
person	9414*
person	9414*
person	9413*
person	9413*
person	9413*
person	9413*
person	9413*
person	9413*
person	9414*

Race:R_0	ZIP code:Z_2
asian	941**
asian	941**
asian	941**
asian	941**
asian	941**
black	941**
black	941**
white	941**
white	941**

Race:R_1	ZIP code:Z_2
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**

Suppression (1)

Race:R_0	ZIP code:Z_0
asian	*94142*
asian	*94141*
asian	94139
asian	94139
asian	94139
black	*94138*
black	*94139*
white	*94139*
white	*94141*

Race:R_1	ZIP code:Z_0
person	*94142*
person	94141
person	94139
person	94139
person	94139
person	*94138*
person	94139
person	94139
person	94141

Race:R_0	ZIP code:Z_1
asian	9414*
asian	9414*
asian	9413*
asian	9413*
asian	9413*
black	9413*
black	9413*
white	*9413**
white	*9414**

Race:R_1	ZIP code:Z_1
person	9414*
person	9414*
person	9413*
person	9413*
person	9413*
person	9413*
person	9413*
person	9413*
person	9414*

Race:R_0	ZIP code:Z_2
asian	941**
asian	941**
asian	941**
asian	941**
asian	941**
black	941**
black	941**
white	941**
white	941**

Race:R_1	ZIP code:Z_2
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**

Suppression (2)

Race:R_0	**ZIP code**:Z_0
asian	94139
asian	94139
asian	94139

Race:R_1	**ZIP code**:Z_0
person	94141
person	94139
person	94139
person	94139
person	94139
person	94139
person	94141

Race:R_0	**ZIP code**:Z_1
asian	9414*
asian	9414*
asian	9413*
asian	9413*
asian	9413*
black	9413*
black	9413*

Race:R_1	**ZIP code**:Z_1
person	9414*
person	9414*
person	9413*
person	9413*
person	9413*
person	9413*
person	9413*
person	9413*
person	9414*

Race:R_0	**ZIP code**:Z_2
asian	941**
asian	941**
asian	941**
asian	941**
asian	941**
black	941**
black	941**
white	941**
white	941**

Race:R_1	**ZIP code**:Z_2
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**

k-Minimal generalization with suppression (1)

Many tables obtained by generalizing attributes and suppressing tuples satisfy *k*-anonymity, but some of them may be either too general or have too many tuples removed.

Obtain *k*-minimal generalization with suppression based on distance vectors in the generalization hierarchy.

k-Minimal generalization with suppression (2)

Criteria for choosing minimal generalization with suppression:

- Minimum absolute distance

  Smallest total number of generalization steps.
- Minimum relative distance

  Smallest total number of relative generalization steps.
- Maximum distribution

  Greatest number of distinct tuples.
- Minimum suppression

  Greatest number of tuples.

Classification of *k*-anonymity techniques (1)

Generalization:

- Attribute level (AG)
- Cell level (CG)

Suppression:

- Tuple level (TS)
- Attribute level (AS)
- Cell level (CS)

Generalization	**Suppression**			
	Tuple	*Attribute*	*Cell*	*None*
Attribute	AG_TS	AG_AS ($\equiv$ AG_)	AG_CS	AG_ ($\equiv$ AG_AS)
Cell	CG_TS	CG_AS	CG_CS ($\equiv$ CG_)	CG_ ($\equiv$ CG_CS)
None	_TS	_AS	_CS	not interesting

Classification of *k*-anonymity techniques (2)

Race	DOB	Sex	ZIP
asian	64/04	F	941**
asian	64/04	F	941**
asian	63/03	M	941**
asian	63/03	M	941**
black	64/09	F	941**
black	64/09	F	941**
white	64/09	F	941**
white	64/09	F	941**

AG_TS

Race	DOB	Sex	ZIP
asian	*	F	*
asian	*	F	*
asian	*	F	*
asian	63/03	M	9413*
asian	63/03	M	9413*
black	64/09	F	9413*
black	64/09	F	9413*
white	64/09	F	*
white	64/09	F	*

AG_CS

Race	DOB	Sex	ZIP
asian	64	F	941**
asian	64	F	941**
asian	64	F	941**
asian	63	M	941**
asian	63	M	941**
black	64	F	941**
black	64	F	941**
white	64	F	941**
white	64	F	941**

AG_≡AG_AS

Race	DOB	Sex	ZIP
asian	64	F	941**
asian	64	F	941**
asian	64	F	941**
asian	63/03	M	94139
asian	63/03	M	94139
black	64/09/27	F	9413*
black	64/09/27	F	9413*
white	64/09/27	F	941**
white	64/09/27	F	941**

CG_≡CG_CS

Race	DOB	Sex	ZIP
asian	*	F	*
asian	*	F	*
asian	*	F	*
asian	*	M	*
asian	*	M	*
black	*	F	*
black	*	F	*
white	*	F	*
white	*	F	*

_AS

Race	DOB	Sex	ZIP
asian	*	F	*
asian	*	F	*
asian	*	F	*
asian	*	M	94139
asian	*	M	94139
*	64/09/27	F	*
*	64/09/27	F	94139
*	64/09/27	F	94139
*	64/09/27	F	*

_CS

k-Anonymity algorithms

The problem of finding minimal k-anonymous tables with attribute generalization and tuple suppression is NP-hard.

Algorithm	**Model**	**Algorithm's type**	**Time complexity**
Samarati	AG_TS	Exact	exponential in $\|QI\|$
Sweeney	AG_TS	Exact	exponential in $\|QI\|$
Bayardo-Agrawal	AG_TS	Exact	exponential in $\|QI\|$
LeFevre-et-al.	AG_TS	Exact	exponential in $\|QI\|$
Aggarwal-et-al. (1)	_CS	$O(k)$-Approximation	$O(kn^2)$
Meyerson-Williams	_CS	$O(k \log k)$-Approximation	$O(n^{2k})$
Aggarwal-et-al. (2)	CG_	$O(k)$-Approximation	$O(kn^2)$
Iyengar	AG_TS	Heuristic	limited number of iterations
Winkler	AG_TS	Heuristic	limited number of iterations
Fung-Wang-Yu	AG_	Heuristic	limited number of iterations

Further studies on *k*-anonymity (1)

- Multidimensional *k*-Anonymity

  Each value can have several possible generalizations.
  Greater time-complexity but better quality.
- *l*-Diversity

  Prevent homogeneity and background knowledge attacks.
  Force *l* different values for sensitive attributes.
- Evaluation of *k*-Anonymity

  Apply data mining techniques to *k*-anonymization results.
- Distributed Algorithms

  Anonymize data distributed through interconnected parties.

Further studies on k-anonymity (2)

- k-Anonymity with Multiple Views

  Data association: two or more attributes are considered more sensitive when their values are associated than when either appear separately.
- k-Anonymity with Micro-Aggregation

  Divide microdata sets into clusters, compute centroids.
- k-Anonymity for Protecting Location Privacy
- k-Anonymity for Communication Protocols