k-Anonymity

Pierangela Samarati
Dipartimento di Tecnologie dell'Informazione
Università degli Studi di Milano
e_mail: samarati@dti.unimi.it

FOSAD 2008

Data protection

- Lots of data!
  - properties are communicated in place of actual identities
  - properties/credentials enrich identities
  - contextual data allow supporting novel scenarios and requirements
- Lots of data communicated, exchanged, shared
- Data need to be protected!

Data collection and disclosure

- Internet provides unprecedented opportunities for the collection and sharing of privacy-sensitive information from and about users
- Information about users is collected every day
- Users have very strong concerns about the privacy of their personal information
- Protecting privacy requires the investigation of many different issues including the problem of protecting released information against inference and linking attacks
  - huge data collections can now be analyzed by powerful techniques (e.g., data mining techniques) and sophisticated algorithms

Statistical data dissemination

Often statistical data (or data for statistical purpose) are released

Such released data can be used to infer information that was not intended for disclosure

Disclosure can:

- occur based on the released data alone
- result from combination of the released data with publicly available information
- be possible only through combination of the released data with detailed external data sources that may or may not be available to the general public

When releasing data, the disclosure risk from the released data should be very low

Macrodata vs microdata

- In the past data were mainly released in tabular form (macrodata) and through statistical databases [CDFS-07b]
- Today many situations require that the specific stored data themselves, called microdata, be released
  - increased flexibility and availability of information for the users
- However microdata are subject to a greater risk of privacy breaches
- The main requirements that must be taken into account are:
  - identity disclosure protection
  - attribute disclosure protection
  - inference channel

Macrodata

Macrodata tables can be classified into the following two groups (types of tables)

- Count/Frequency. Each cell of the table contains the number of respondents (count) or the percentage of respondents (frequency) that have the same value over all attributes of analysis associated with the table
- Magnitude data. Each cell of the table contains an aggregate value of a *quantity of interest* over all attributes of analysis associated with the table

Count table – Example

Two-dimensional table showing the number of beneficiaries by county and size of benefit

	Benefit						
County	**$0-19**	**$20-39**	**$40-59**	**$60-79**	**$80-99**	**$100+**	**Total**
A	2	4	18	20	7	1	52
B	-	-	7	9	-	-	16
C	-	6	30	15	4	-	55
D	-	-	2	-	-	-	2

Magnitude table – Example

Average number of days spent in the hospital by respondents with a disease

	Hypertension	**Obesity**	**Chest Pain**	**Short Breath**	**Tot**
M	2	8.5	23.5	3	37
F	3	30.5	0	5	38.5
Tot	5	39	23.5	8	75.5

Microdata table – Example

Records about delinquent children in county Alfa

N	Child	County	Educ. HH	Salary HH	Race HH
1	John	Alfa	very high	201	black
2	Jim	Alfa	high	103	white
3	Sue	Alfa	high	77	black
4	Pete	Alfa	high	61	white
5	Ramesh	Alfa	medium	72	white
6	Dante	Alfa	low	103	white
7	Virgil	Alfa	low	91	black
8	Wanda	Alfa	low	84	white
9	Stan	Alfa	low	75	white
10	Irmi	Alfa	low	62	black
11	Renee	Alfa	low	58	white
12	Virginia	Alfa	low	56	black
13	Mary	Alfa	low	54	black
14	Kim	Alfa	low	52	white
15	Tom	Alfa	low	55	black
16	Ken	Alfa	low	48	white
17	Mike	Alfa	low	48	white
18	Joe	Alfa	low	41	black
19	Jeff	Alfa	low	44	black
20	Nancy	Alfa	low	37	white

Information disclosure

Several different definitions of disclosure and different types of disclosure have been proposed

Disclosure relates to improper attribution of information to a respondent, whether an individual or an organization.

There is disclosure when:

- a respondent is identified from released data (identity disclosure)
- sensitive information about a respondent is revealed through the released data (attribute disclosure)
- the released data make it possible to determine the value of some characteristic of a respondent more accurately than otherwise would have been possible (inferential disclosure)

Identity disclosure

It occurs if a third party can identify a subject or respondent from the released data

Revealing that an individual is a respondent or subject of a data collection may or may not violate confidentiality requirements

- Macrodata: revealing identity is generally not a problem, unless the identification leads to divulging confidential information (attribute disclosure)
- Microdata: identification is generally regarded as a problem, since microdata records are detailed; identity disclosure usually implies in this case also attribute disclosure

Attribute disclosure

It occurs when confidential information about a respondent is revealed and can be attributed to it

It may occur when confidential information is revealed exactly or when it can be closely estimated

It comprises identification of the respondent and divulging confidential information pertaining to the respondent

Inferential disclosure

It occurs when information can be inferred with high confidence from statistical properties of the released data

E.g., the data may show a high correlation between income and purchase price of home. As purchase price of home is typically public information, a third party might use this information to infer the income of a respondent

It is difficult to take into consideration this type of disclosure for two reasons

- if disclosure is equivalent to inference, no data could be released
- inferences are designed to predict aggregate behavior, not individual attributes, and are then often poor predictors of individual data values

Restricted data and restricted access (1)

The choice of statistical disclosure limitation methods depends on the nature of the data products whose confidentiality must be protected

Some microdata include explicit identifiers (e.g., name, address, or Social Security number)

Removing such identifiers is a first step in preparing for the release of microdata for which the confidentiality of individual information must be protected

Restricted data and restricted access (2)

The confidentiality of individual information can be protected by:

- restricting the amount of information in released tables and microdata (restricted data)
- imposing conditions on access to the data products (restricted access)
- some combination of these two strategies

Disclosure protection techniques

The protection techniques include:

- sampling: data confidentiality is protected by conducting a sample survey rather than a census
- special rules: designed for specific tables, they impose restrictions on the level of detail that can be provided in a table
- threshold rule: rules that to protect sensitive cells
  - cell suppression
  - random rounding
  - controlled rounding
  - confidentiality edit

The anonymity problem

- The amount of privately owned records that describe each citizen's finances, interests, and demographics is increasing every day
- These data are de-identified before release, that is, any explicit identifier (e.g., SSN) is removed
- De-identification is not sufficient
- Most municipalities sell population registers that include the identities of individuals along with basic demographics
- These data can then be used for linking identities with de-identified information $\Rightarrow$**re-identification**

Re-identification

In 2000, the US population was uniquely identifiable by:

- year of birth, 5-digit ZIP code: 0,2%
- year of birth, county: 0,0%
- year and month of birth, 5-digit ZIP code: 4,2%
- year and month of birth, county: 0,2%
- year, month, and day of birth, 5-digit ZIP code: 63,3%
- year, month, and day of birth, county: 14,8%

The anonymity problem: Example

SSN	Name	Race	Date of birth	Sex	ZIP	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	DOB	Sex	Status
. . . . .	. . . . .	. . . . .	. . . . .	. . . . .	. . . . .	. . . . .
. . . . .	. . . . .	. . . . .	. . . . .	. . . . .	. . . . .	. . . . .
Sue J. Doe	900 Market St.	San Francisco	*94142*	*64/04/12*	*F*	*divorced*
. . . . .	. . . . .	. . . . .	. . . . .	. . . . .	. . . . .	. . . . .

k-anonymity (1)

- k-anonymity, together with its enforcement via generalization and suppression, has been proposed as an approach to protect respondents'identities while releasing truthful information [S-01]
- k-anonymity tries to capture the following requirement:

  the released data should be indistinguishably related to no less than a certain number of respondents

- Quasi-identifier: Set of attributes that can be exploited for linking (whose release must be controlled)

k-anonymity (2)

- The basic idea is therefore to translate the above-mentioned requirement into a requirement on the released data

Each release of data must be such that every combination of values of quasi-identifiers can be indistinctly matched to at least k respondents

- k-anonymity requires that, in the released table itself, the respondents be indistinguishable (within a given set) with respect to a set of attributes
- To guarantee the k-anonymity requirement, k-anonymity requires each quasi-identifier value in the released table to have at least k occurrences
  - sufficient condition for the k-anonymity requirement

Generalization and suppression

- Generalization. The values of a given attribute by using more general values. Based on the definition of a generalization hierarchy
  - E.g., consider attribute ZIP code and suppose that a step in the corresponding generalization hierarchy consists in suppressing the least significant digit in the ZIP code
  - If we choose to apply one generalization step, values 20222, 20223, 20238, and 20239 are generalized to 2022* and 2023*.
- Suppression. It is a well-known technique that consists in protecting sensitive information by removing it.
  - The introduction of suppression can reduce the amount of generalization necessary to satisfy the k-anonymity constraint.

Domain generalization hierarchy

- DGH$_D$ = (Dom, $\leq_D$):

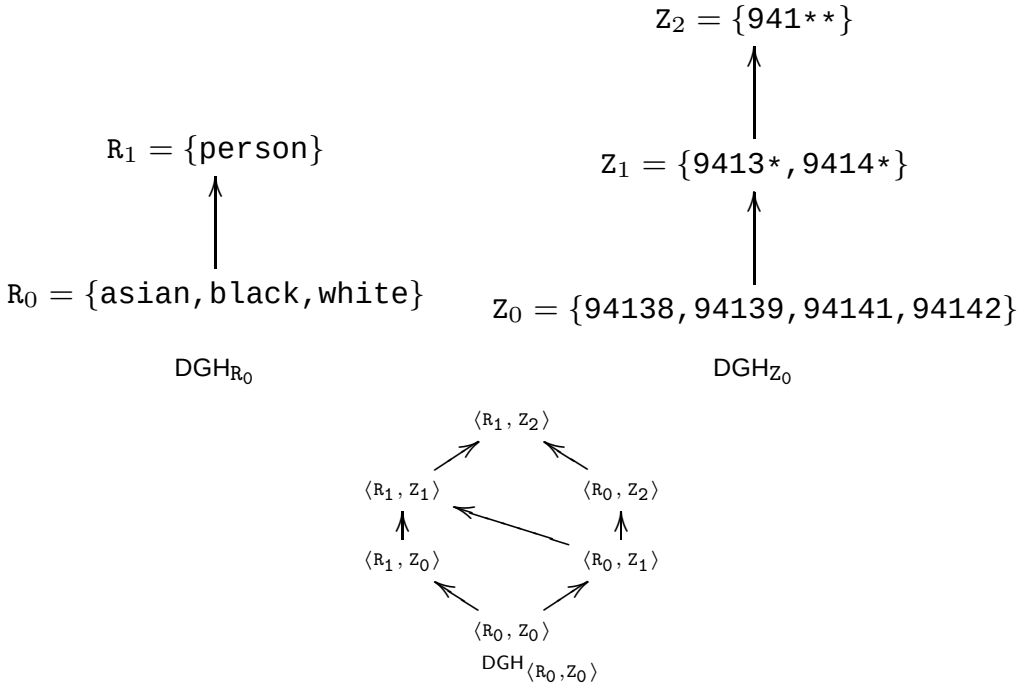
  C1: $\forall D_i, D_j, D_z \in$ Dom:

  $$D_i \leq_D D_j, D_i \leq_D D_z \Rightarrow D_j \leq_D D_z \vee D_z \leq_D D_j$$

  C2: all maximal elements of Dom are singleton.

- Given a domain tuple $DT = \langle D_1, \ldots, D_n \rangle$ such that $D_i \in$ Dom, $i = 1, \ldots, n$, the domain generalization hierarchy of DT is $\text{DGH}_{DT} = \text{DGH}_{D_1} \times \ldots \times \text{DGH}_{D_n}$

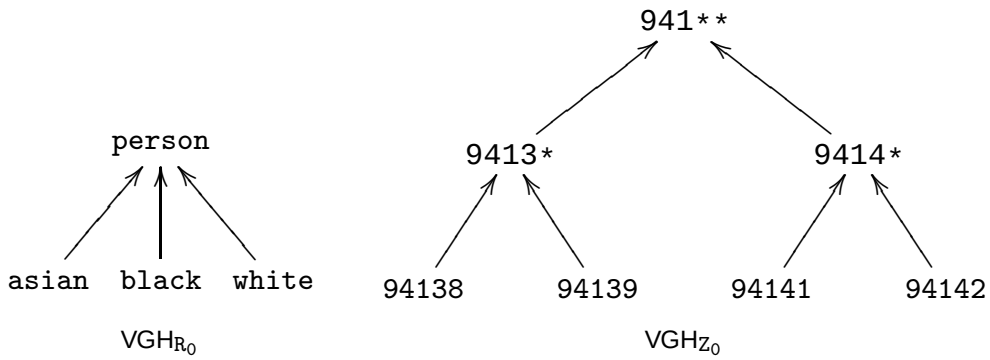
Examples of domain generalization hierarchies



Value generalization hierarchy

- A value generalization relationship, denoted $\leq_V$, associates with each value in domain D_i a unique value in domain D_j, direct generalization of D_i
- The value generalization relationship implies the existence, for each domain D, of a value generalization hierarchy, denoted VGH$_D$
- VGH$_D$ is a tree, where the leaves are the values in D and the root (i.e., the most general value) is the value in the maximum element in DGH$_D$

Examples of value generalization hierarchies



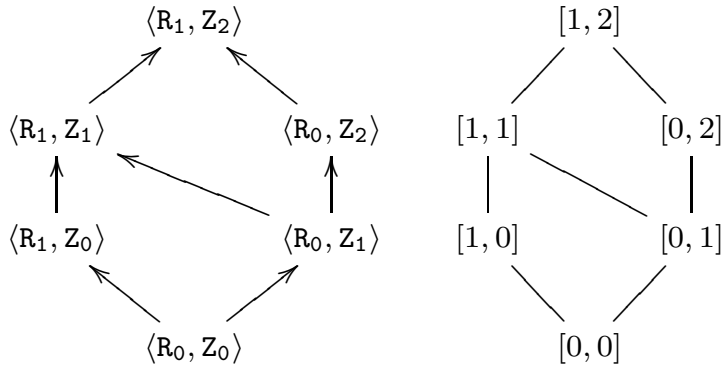
Generalized table with suppression

Let T_i and T_j be two tables defined on the same set of attributes. Table T_j is said to be a *generalization (with tuple suppression)* of table T_i, denoted $T_i \preceq T_j$, if:

1. $|T_j| \leq |T_i|$;
2. the domain $dom(A, T_j)$ of each attribute A in T_j is equal to, or a generalization of, the domain $dom(A, T_i)$ of attribute A in T_i;
3. it is possible to define an injective function associating each tuple t_j in T_j with a tuple t_i in T_i, such that the value of each attribute in t_j is equal to, or a generalization of, the value of the corresponding attribute in t_i.

k-minimal generalization with suppression (1)

- Distance vector. Let $T_i(A_1, \ldots, A_n)$ and $T_j(A_1, \ldots, A_n)$ be two tables such that $T_i \preceq T_j$. The distance vector of T_j from T_i is the vector $DV_{i,j} = [d_1, \ldots, d_n]$, where each d_z, $z = 1, \ldots, n$, is the length of the *unique* path between dom(A_z, T_i) and dom(A_z, T_j) in the domain generalization hierarchy DGH_{D_z}.



k-minimal generalization with suppression (2)

Let T_i and T_j be two tables such that $T_i \preceq T_j$, and let MaxSup be the specified threshold of acceptable suppression. T_j is said to be a *k-minimal* generalization of table T_i iff:

1. T_j satisfies k-anonymity enforcing minimal required suppression, that is, T_j satisfies k-anonymity and $\forall T_z : T_i \preceq T_z, DV_{i,z} = DV_{i,j}, T_z$ satisfies k-anonymity $\Rightarrow |T_j| \geq |T_z|$
2. $|T_i| - |T_j| \leq$ MaxSup
3. $\forall T_z : T_i \preceq T_z$ and T_z satisfies conditions 1 and 2 $\Rightarrow \neg(DV_{i,z} < DV_{i,j})$.

Examples of 2-minimal generalization

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

PT

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

$GT_{[1,0]}$

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

$GT_{[0,1]}$

Computing a preferred generalization

Different *preference criteria* can be applied in choosing a preferred minimal generalization, among which:

- minimum absolute distance prefers the generalization(s) with the smallest absolute distance, that is, with the smallest total number of generalization steps (regardless of the hierarchies on which they have been taken);
- minimum relative distance prefers the generalization(s) with the smallest relative distance, that is, that minimizes the total number of relative steps (a step is made relative by dividing it over the height of the domain hierarchy to which it refers);
- maximum distribution prefers the generalization(s) with the greatest number of distinct tuples;
- minimum suppression prefers the generalization(s) that suppresses less tuples, that is, the one with the greatest cardinality.

Classification of k-Anonymity techniques (1)

Generalization and suppression can be applied at different levels of granularity.

- Generalization can be applied at the level of single column (i.e., a generalization step generalizes all the values in the column) or single cell (i.e., for a specific column, the table may contain values at different generalization levels)
- Suppression can be applied at the level of row (i.e., a suppression operation removes a whole tuple), column (i.e., a suppression operation obscures all the values of a column), or single cells (i.e., a k-anonymized table may wipe out only certain cells of a given tuple/attribute)

Classification of k-Anonymity techniques (2)

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

2-anonymized tables wrt different models (1)

Race	DOB	Sex	ZIP
asian	64/04/12	F	94142
asian	64/09/13	F	94141
asian	64/04/15	F	94139
asian	63/03/13	M	94139
asian	63/03/18	M	94139
black	64/09/27	F	94138
black	64/09/27	F	94139
white	64/09/27	F	94139
white	64/09/27	F	94141

(**a**) PT

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**

(**b**) **AG_TS**

2-anonymized tables wrt different models (2)

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	*

(c) **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**

(d) **AG_$\equiv$AG_AS**

2-anonymized tables wrt different models (3)

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**

(e) **CG_$\equiv$CG_CS**

Race	DOB	Sex	ZIP

(f) **_TS**

2-anonymized tables wrt different models (4)

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

(**g**) **_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	*

(**h**) **_CS**

Algorithms for computing a k-anonymous table

- The problem of finding minimal k-anonymous tables, with attribute generalization and tuple suppression, is computationally hard
- The majority of the exact algorithms proposed in literature have computational time exponential in the number of the attributes composing the quasi-identifier
  - when the number $|QI|$ of attributes in the quasi-identifier is small compared with the number n of tuples in the private table PT, these exact algorithms with attribute generalization and tuple suppression are practical
- Recently many exact algorithms for producing k-anonymous tables through attribute generalization and tuple suppression have been proposed

Algorithms for AG_TS and AG_

Samarati's algorithm (1) [S-01]

- Each path in DGH$_{DT}$ represents a generalization strategy for PT
- We call locally minimal generalization the lowest node of each path satisfying k-anonymity
- Properties exploited by the algorithm:
  1. each k-minimal generalization is locally minimal with respect to a path (but the converse is not true)
  2. going up in the hierarchy the number of tuples that must be removed to guarantee k-anonymity decreases
- If there is no solution that guarantees k-anonymity suppressing less than MaxSup tuples at height h, there cannot exist a solution, with height lower than h that guarantees it

Samarati's algorithm (2)

- The algorithm adopts a *binary search* on the lattice of distance vectors:
  1. evaluate solutions at height $\lfloor h/2 \rfloor$
  2. if there exists at least a solution satisfying k-anonymity
     - then evaluates solutions at height $\lfloor h/4 \rfloor$
     - otherwise evaluates solutions at height $\lfloor 3h/4 \rfloor$
  3. until the algorithm reaches the lowest height for which there is a distance vector that satisfies k-anonymity
- To reduce the computational cost, it adopts a distance vector matrix that avoids the explicit computation of each generalized table

Samarati's algorithm – Example (1)

Suppose $k = 2$ and MaxSup=2.

Compute first solutions at height 1: $\mathsf{GT}_{[1,0]}$ and $\mathsf{GT}_{[0,1]}$

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

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

Both the generalized tables satisfy 2-anonymity

Samarati's algorithm – Example (2)

Compute solutions at height 0: $GT_{[0,0]}$

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

The generalized table does not satisfy 2-anonymity

Samarati's algorithm – Example (3)

Distance vector matrix for the considered table

	t_1	t_2	$t_3/t_4/t_5$	t_6	t_7	t_8	t_9
t_1	$[0,0]$	$[0,1]$	$[0,2]$	$[1,2]$	$[1,2]$	$[1,2]$	$[1,1]$
t_2	$[0,1]$	$[0,0]$	$[0,2]$	$[1,2]$	$[1,2]$	$[1,2]$	$[1,0]$
t_6	$[1,2]$	$[1,2]$	$[1,1]$	$[0,0]$	$[0,1]$	$[1,1]$	$[1,2]$
t_7	$[1,2]$	$[1,2]$	$[1,0]$	$[0,1]$	$[0,0]$	$[1,0]$	$[1,2]$
t_8	$[1,2]$	$[1,2]$	$[1,0]$	$[1,1]$	$[1,0]$	$[0,0]$	$[0,2]$
t_9	$[1,1]$	$[1,0]$	$[1,2]$	$[1,2]$	$[1,2]$	$[0,2]$	$[0,0]$

k-Optimize algorithm (1) [BA-05]

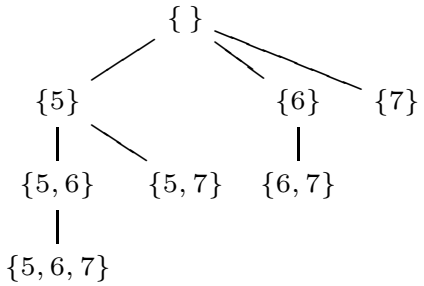
- Order attributes in *QI* and the values in their domains
- Associate an integer *index* value with each domain value, following the defined order

Race			ZIP			
⟨[asian]	[black]	[white]⟩	⟨[94138]	[94139]	[94141]	[94142]⟩
1	2	3	4	5	6	7

- A generalization is the union of individual index values
- The least value in an attribute domain is omitted. E.g., $\{6\}$ corresponds to:
  - Race: $\{1\}$, that is: ⟨[asian or black or white]⟩
  - ZIP: $\{4, 6\}$, that is: ⟨[94138 or 94139],[94141 or 94142]⟩
- Order of values within domains has impact on generalization

k-Optimize algorithm (2)

- k-Optimize builds a set enumeration tree over the set I of indexes



- The root node of the tree is the empty set
- The children of n are the sets obtained by appending a single element i of I to n, such that $\forall i' \in n, i > i'$
- k-Optimize visits the tree computing the cost of each node and keeping the best found
- The algorithm *prunes* subtrees by evaluating their lower bounds

Incognito algorithm [LDR-05]

k-anonymity with respect to a proper subset of QI is a necessary (not sufficient) condition for k-anonymity with respect to QI

- Iteration 1: check k-anonymity for each attribute in QI, discarding generalizations that do not satisfy k-anonymity
- Iteration 2: combine the remaining generalizations in pairs and check k-anonymity for each couple obtained

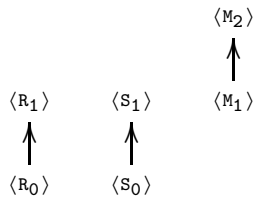
  ...
- Iteration i: consider all the i-uples of attributes, obtained combining generalizations that satisfied k-anonymity at iteration $i-1$. Discard non k-anonymous solutions

  ...
- Iteration $|QI|$ returns the final result

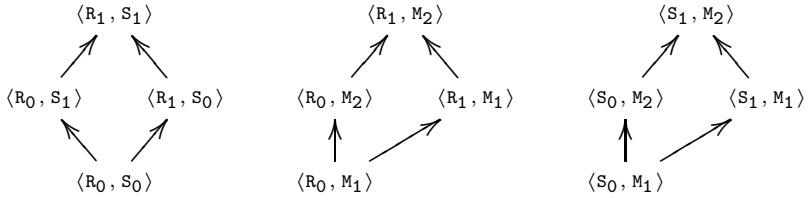
Incognito adopts a bottom-up approach for the visit of DGHs

Incognito: Example (1)

Iteration 1

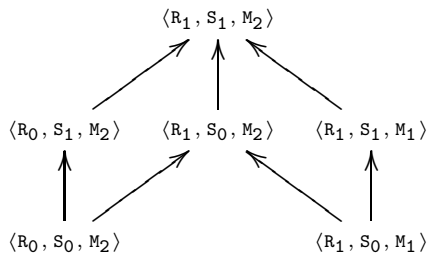


Iteration 2



Incognito: Example (2)

Iteration 3



Heuristic algorithms

- The exact algorithms have complexity exponential in the size of QI
- Heuristic algorithms have been proposed
  - [I-02]: based on genetic algorithms, it solves the k-anonymity problem using an incomplete stochastic search method
  - [W-04]: based on simulated annealing for finding locally minimal solutions, it requires high computational time and does not assure the quality of the solution
  - [FWY-05]: top-down heuristic to make a table to be released k-anomymous; it starts from the most general solution, and iteratively specializes some values of the current solution until the k-anonymity requirement is violated
- No bounds on efficiency and goodness of the solutions can be given
- Experimental results can be used to assess the quality of the solution retrieved

Algorithms for _CS and CG_

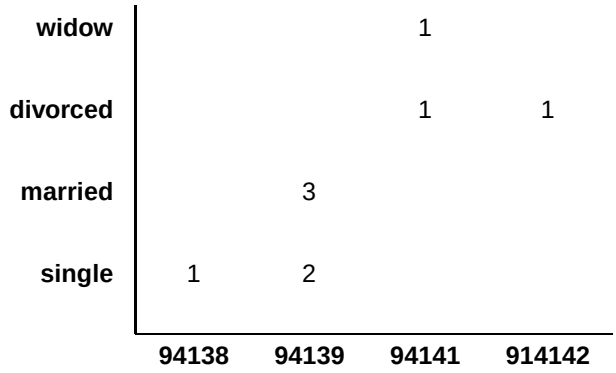
Mondrian multidimensional algorithm [LDR-06]

- Each attribute in QI represents a dimension
- Each tuple in PT represents a point in the space defined by QI
- Tuples with the same QI value are represented by associating the numbers of occurrences with points
- The multi-dimensional space is partitioned by splitting dimensions such that each area contains at least k occurrences of point values
- All the points in a region are generalized to a unique value
- The corresponding tuples are substituted by the computed generalization

Mondrian multidimensional algorithm – Example (1)

Private table

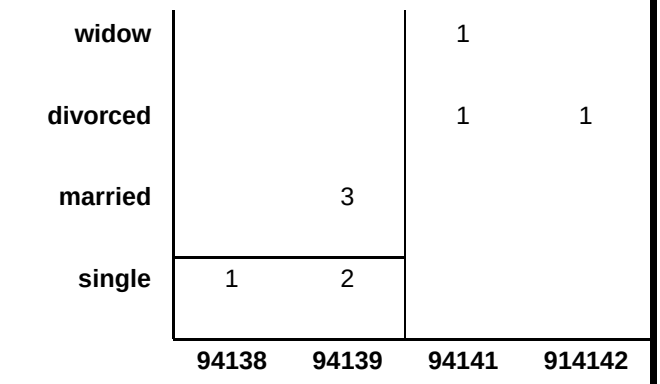
Marital status	ZIP
divorced	94142
divorced	94141
married	94139
married	94139
married	94139
single	94138
single	94139
single	94139
widow	94141



Mondrian multidimensional algorithm – Example (2)

3-anonymous table

Marital status	ZIP
divorced or widow	9414*
divorced or widow	9414*
married	94139
married	94139
married	94139
single	9413*
single	9413*
single	9413*
divorced or widow	9414*



Approximation algorithms

- Approximation algorithms for general and specific values of k (e.g., 1.5-approximation for 2-anonymity, and 2-approximation for 3-anonymity [AFKMPTZ-05b])
- Approximation algorithm for **_CS**
  - [MW-04]: $O(k \log(k))$-approximation
  - [AFKMPTZ-05a]: with unbounded value of k, $O(k)$-approximation solution
- Approximation algorithm for **CG_**
  - [AFKMPTZ-05b]: with unbounded value of k, $O(k)$-approximation solution

k-anonymity revisited [GMT-08]

- In the case of cell generalization (CG) there is no need to require presence of k equal tuples to guarantee k-anonymity
- Require existence of k corresponding tuples in T (PT, resp.) for each tuple in PT (T, resp.)

k-anonymity revisited – Example

Race	ZIP
White	94138
Black	94139
Asian	94141
Asian	94141
Asian	94142

PT

Race	ZIP
Person	9413*
Person	9413*
Asian	9414*
Asian	9414*
Asian	9414*

2-anonymity

Race	ZIP
Person	9413*
Person	9413*
Asian	9414*
Asian	9414*
Asian	9414*

2-anonymity (revisited)

Race	ZIP
Person	9413*
Person	9413*
Asian	9414*
Asian	9414*
Asian	94142

Race	ZIP
Person	9413*
Person	9413*
Asian	94141
Asian	94141
Asian	9414*

no 2-anonymity

Attribute Disclosure

2-anonymous table according to the AG_ model

k-anonymity is vulnerable to some attacks [MGK-06,S-01]

Race	DOB	Sex	ZIP	Disease
asian	64	F	941**	hypertension
asian	64	F	941**	obesity
asian	64	F	941**	chest pain
asian	63	M	941**	obesity
asian	63	M	941**	obesity
black	64	F	941**	short breath
black	64	F	941**	short breath
white	64	F	941**	chest pain
white	64	F	941**	short breath

Homogeneity attack

- All tuples with a quasi-identifier value in a k-anonymous table have the same sensitive attribute value
  - an attacker knows that `Carol` is a `black female` and that her data are in the microdata table
  - the attacker can infer that `Carol` suffers of `short breath`

Race	DOB	Sex	ZIP	Disease
...	...	...	...	...
black	64	F	941**	short breath
black	64	F	941**	short breath
...	...	...	...	...

Background knowledge attack

- Based on a prior knowledge of some additional external information. E.g.
  - an attacker knows that `Hellen` is a `white female`
  - the attacker can infer that the disease of `Hellen` is either `chest pain` or `short breath`
  - If the attacker knows that the `Hellen` runs 2 hours a day, she can infer that `Hellen`'s disease is `chest pain`

Race	DOB	Sex	ZIP	Disease
...	...	...	...	...
white	64	F	941**	chest pain
white	64	F	941**	short breath

ℓ-diversity [MGK-06]

- A q-block (i.e., set of tuples with the same value for QI) in T is ℓ-diverse if it contains at least ℓ different values for the sensitive attribute in T
  $\Rightarrow$ the homogeneity attack is not possible anymore
  $\Rightarrow$ the background knowledge attack becomes more complicate
- T is ℓ-diverse if all its q-blocks are *ℓ-diverse*
- ℓ-diversity is monotonic with respect to the generalization hierarchies considered for k-anonymity purposes
- Any algorithm for k-anonymity can be extended to enforce the ℓ-diverse property

Skewness attack

ℓ-diversity leaves space to attacks based on the distribution of values inside q-blocks

- skewness attack: happens when the distribution in a q-block is different than in the original population
- Suppose that 20% of the population suffer from diabetes:

Race	DOB	Sex	ZIP	Disease
black	64	F	941**	diabetes
black	64	F	941**	short breath
black	64	F	941**	diabetes
black	64	F	941**	diabetes

Similarity attack

- similarity attack: happens when a q-block has different but semantically similar values for the sensitive attribute

Race	DOB	Sex	ZIP	Disease
black	64	F	941**	stomach ulcer
black	64	F	941**	stomach ulcer
black	64	F	941**	gastritis

t-closeness [LLV-07]

- A q-block respects t-closeness if the distance between the distribution of the values of the sensitive attribute in the q-block and in the considered population is lower than t
- T respects t-closeness if all its q-blocks respect t-closeness
- t-closeness is monotonic with respect to the generalization hierarchies considered for k-anonymity purposes
- Any algorithm for k-anonymity can be extended to enforce the t-closeness property

Still a lot to be done ... (1)

- Extensions and enrichment of the definition. k-anonymity captures only the defence against identity disclosure attacks; it is exposed to attribute disclosure attacks (some approaches but research is still to be done)
- Protection against utility measures. Research is needed to develop measures to allow users to assess, besides the protection offered by the data, the utility of the released data
- Efficient algorithms. Computing a table that satisfies k-anonymity guaranteeing minimality is an NP-hard problem and therefore computationally expensive
- New techniques. The k-anonymity property is not tied to a specific technique and alternative techniques could be investigated

Still a lot to be done ... (2)

- Merging of different tables and views. The original k-anonymity proposal as well as most subsequent work assume:
  - the existence of a single table to be released
  - the released table contains at most one tuple for each respondents

  Work is needed to release these two constraints.
- External knowledge. k-anonymity did not model external knowledge that can be further exploited for inference and expose the data to identity or attribute disclosure.

References (1)

- [AFKMPTZ-05a] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, A. Zhu, "Anonymizing tables," in *Proc. of the 10th International Conference on Database Theory (ICDT'05)*, Edinburgh, Scotland, 2005.
- [AFKMPTZ-05b] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, A. Zhu, "Approximation algorithms for k-anonymity," *Journal of Privacy Technology*, paper number 20051120001.
- [BA-05] R.J. Bayardo, R. Agrawal, "Data privacy through optimal k-anonymization," in *Proc. of the 21st International Conference on Data Engineering (ICDE'05)*, pp. 217–228, Tokyo, Japan, 2005.
- [BMW-07] C. Bettini, S. Mascetti, X.S. Wang, "Privacy protection through anonymity in location-based services," in *Digital Privacy: Theory, Technologies, and Practices*, Taylor and Francis, 2007.
- [CDFS-07a] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, P. Samarati, "k-Anonymity," in *Secure Data Management in Decentralized Systems*, T. Yu and S. Jajodia (eds), Springer-Verlag, 2007.

References (2)

- [CDFS-07b] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, P. Samarati, "Microdata protection," in *Secure Data Management in Decentralized Systems*, T. Yu and S. Jajodia (eds), Springer-Verlag, 2007.
- [FWY-05] B. Fung, K. Wang, P. Yu, "Top-down specialization for information and privacy preservation," in *Proc. of the 21st International Conference on Data Engineering (ICDE'05)*, Tokyo, Japan, 2005.
- [GMT-08] A. Gionis, A. Mazza and T. Tassa, "k-Anonymization revisited," in *Proc. of the International Conference on Data Engineering,* Cancun, Mexico, 2008.
- [YWJ-05] S. Jajodia, C. Yao, X.S. Wang, "Checking for k-anonymity violation by views," in *Proc. of the 31st International Conference on Very Large Data Bases (VLDB'05)*, Trondheim, Norway, 2005.
- [LDR-05] K. LeFevre, D.J. DeWitt, R. Ramakrishnan, "Incognito: Efficient fulldomain k-anonymity," in *Proc. of the 24th ACM SIGMOD International Conference on Management of Data*, pp. 4960, Baltimore, Maryland, USA, 2005.
- [LDR-06] K. LeFevre, D.J. DeWitt, R. Ramakrishnan, "Mondrian multidimensional k-anonymity," in *Proc. of 22nd International Conference on Data Engineering*, Atlanta, GA, USA, 2006.

References (3)

- [LLV-07] N. Li, T. Li, and S. Venkatasubramanian. "t-closeness: Privacy beyond k-anonymity and ℓ-diversity," In *Proc. of the 23nd International Conference on Data Engineering*, Istanbul, Turkey, April 2007.
- [MGK-06] A. Machanavajjhala, J. Gehrke, D. Kifer "ℓ-diversity: Privacy beyond k-anonymity," in *Proc. of the International Conference on Data Engineering (ICDE'06)*, Atlanta, GA, USA, 2006.
- [MBWJ-06] S. Mascetti, C. Bettini, X.S. Wang, S. Jajodia, "k-Anonymity in databases with timestamped data," in *Proc. of 13th International Symposium on Temporal Representation and Reasoning*, 2006.
- [MW-04] A. Meyerson and R. Williams, "On the complexity of optimal k-anonymity," in *Proc. of the 23rd ACM-SIGMOD-SIGACT-SIGART Symposium on the Principles of Database Systems*, Paris, France, 2004.
- [S-01] P. Samarati, "Protecting respondents' identities in microdata release," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, n. 6, November/December 2001, pp. 1010-1027.