Protecting Privacy in Data Publication

Pierangela Samarati
Dipartimento di Informatica
Università degli Studi di Milano
pierangela.samarati@unimi.it

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 different issues, including the problem of protecting released information against inference and linking attacks which are becoming easier and easier because of the increased information availability and ease of access

Statistical DBMS vs statistical data

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

- statistical DBMS [AW-89]
  - the DBMS responds only to statistical queries
  - need run time checking to control information (indirectly) released
- statistical data [CDFS-07]
  - publish statistics
  - control on indirect release performed before publication

Disclosure risk

Statistical data, even if 'anonymized', 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 of sensitive information should be very low

Macrodata vs microdata

- In the past data were mainly released in tabular form (macrodata) and through statistical databases
- Today many situations require that the specific stored data themselves, called microdata, be released
  - increased flexibility and availability of information for the users
- 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
  - inferential disclosure protection

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

Disclosure relates to attribution of sensitive information to a respondent (an individual or 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 characteristics of a respondent even if no released record refers to the respondent (inferential disclosure)

Identity disclosure

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

Revealing that an individual is a respondent in 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 her

Confidential information may be:

- revealed exactly
- closely estimated

Inferential disclosure

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

EXAMPLE: 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

Inference disclosure does not always represent a risk:

- statistical data are released for enabling users to infer and understand relationships between variables

- 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)

Confidentiality can be protected by:

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

Disclosure protection techniques for macrodata

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 protect sensitive cells, for instance:
  - cell suppression
  - random rounding
  - controlled rounding
  - confidentiality edit

Disclosure protection techniques for microdata

The classical protection techniques (often applied to protect microdata before computing statistics) can be classified as follows:

- masking techniques: transform the original set of data by not releasing or perturbing their values
  - non-perturbative: the original data are not modified, but some data are suppressed and/or some details are removed (e.g., sampling, local suppression, generalization)
  - perturbative: the original data are modified (e.g., rounding, swapping)
- synthetic data generation techniques: release plausible but synthetic values instead of the real ones
  - fully synthetic: the released dataset contains synthetic data only
  - partially synthetic: the released dataset contains a mix of original and synthetic data

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 $\Longrightarrow$**re-identification**

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
.....	.....	.....	.....	.....	.....	.....

Classification of attributes in a microdata table

The attributes in the original microdata table can be classified as:

- identifiers: attributes that uniquely identify a microdata respondent (e.g., SSN uniquely identifies the person with which is associated)
- quasi-identifiers: attributes that, in combination, can be linked with external information to reidentify all or some of the respondents to whom information refers or reduce the uncertainty over their identities (e.g., DoB, ZIP, and Sex)
- confidential: attributes of the microdata table that contain sensitive information (e.g., Disease)
- non confidential: attributes that the respondents do not consider sensitive and whose release does not cause disclosure

Re-identification

A study of the 2000 census data reported that 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%

Factors contributing to disclosure risk (1)

Possible sources of the disclosure risk of microdata

- Existence of high visibility records. Some records on the file may represent respondents with unique characteristics such as very unusual jobs (e.g., movie star) or very large incomes

- Possibility of matching the microdata with external information. There may be individuals in the population who possess a unique or peculiar combination of the characteristic variables on the microdata
  - if some of those individuals happen to be chosen in the sample of the population, there is a disclosure risk
  - note that the identity of the individuals that have been chosen should be kept secret

Factors contributing to disclosure risk (2)

The possibility of linking or its precision increases with:

- the existence of a high number of common attributes between the microdata table and the external sources
- the accuracy or resolution of the data
- the number of outside sources, not all of which may be known to the agency releasing the microdata

Factors contributing to decrease the disclosure risk (1)

- A microdata table often contains a subset of the whole population
  - this implies that the information of a specific respondent, which a malicious user may want to know, may not be included in the microdata table

- The information specified in microdata tables released to the public is not always up-to-date (often at least one or two-year old)
  - the values of the attributes of the corresponding respondents may have changed in the meanwhile
  - the age of the external sources of information used for linking may be different from the age of the information contained in the microdata table

Factors contributing to decrease the disclosure risk (2)

- A microdata table and the external sources of information naturally contain noise that decreases the ability to link the information
- A microdata table and the external sources of information can contain data expressed in different forms thus decreasing the ability to link information

Measures of risk

Measuring the disclosure risk requires considering

- the probability that the respondent for whom an intruder is looking is represented on both the microdata and some external file
- the probability that the matching variables are recorded in a linkable way on the microdata and on the external file
- the probability that the respondent for whom the intruder is looking is unique (or peculiar) in the population of the external file

The percentage of records representing respondents who are unique in the population (population unique) plays a major role in the disclosure risk of microdata (with respect to the specific respondent)

Note that each population unique is a sample unique; the vice-versa is not true

k-anonymity [S-01] (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

- 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)

- Basic idea: translate the k-anonymity 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
- In the released table the respondents must be indistinguishable (within a given set) with respect to a set of attributes
- k-anonymity requires that each quasi-identifier value appearing in the released table must have at least k occurrences
  - sufficient condition for the satisfaction of k-anonymity requirement

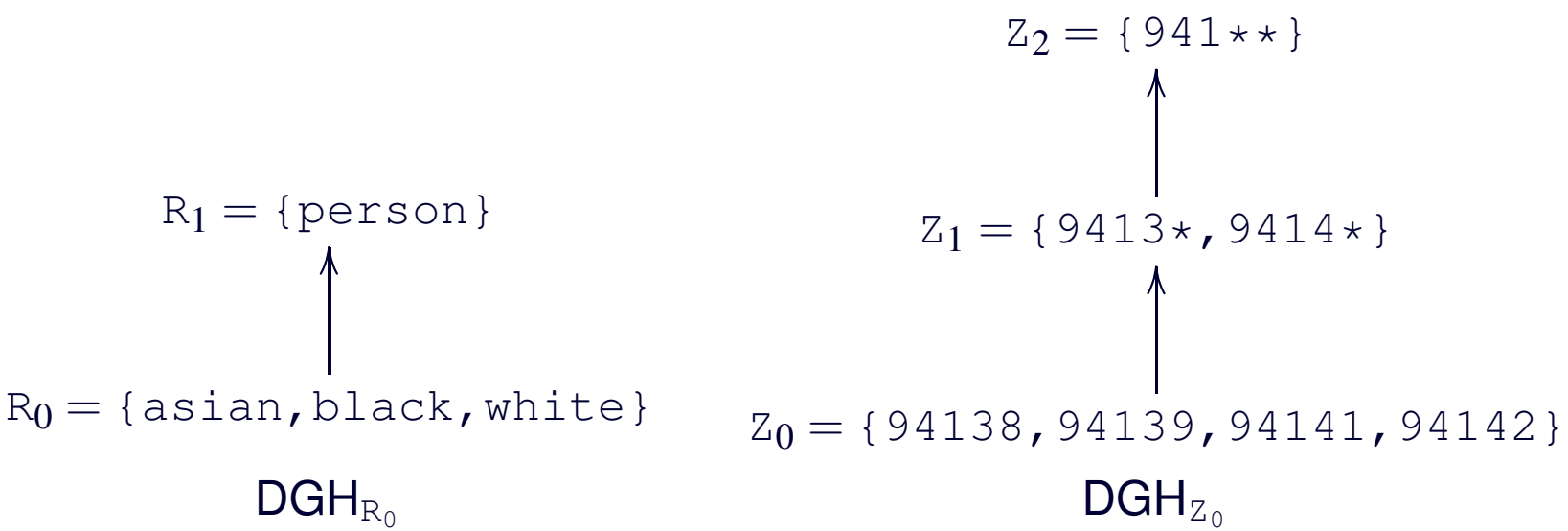
Generalization and suppression

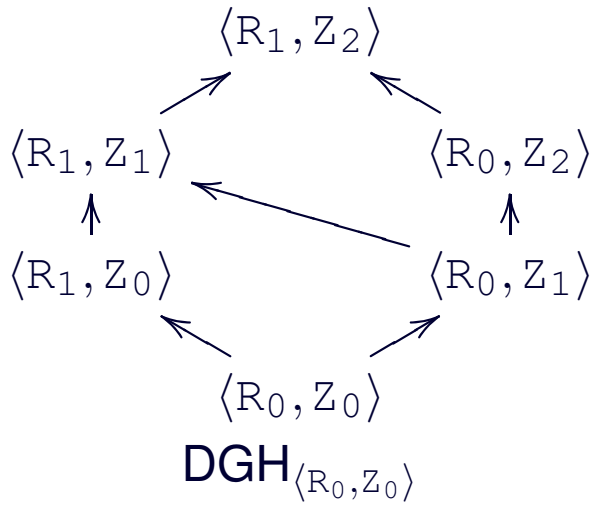
- Generalization. The values of a given attribute are substituted by using more general values. Based on the definition of a generalization hierarchy
  - Example: consider attribute ZIP code and suppose that a step in the corresponding generalization hierarchy consists in suppressing the least significant digit in the ZIP code
    With one generalization step: 20222 and 20223 become 2022*; 20238 and 20239 become 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

- A generalization relationship $\leq_D$ defines a mapping between domain D and its generalizations
- Given two domains $D_i, D_j \in$ Dom, $D_i \leq_D D_j$ states that the values in domain D_j are generalizations of values in D_i
- $\leq_D$ implies the existence, for each domain D, of a domain generalization hierarchy $\text{DGH}_D = (\text{Dom}, \leq_D)$:
  - $\forall D_i, D_j, D_z \in$ Dom:
    $D_i \leq_D D_j, D_i \leq_D D_z \Longrightarrow D_j \leq_D D_z \vee D_z \leq_D D_j$
  - 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 $\text{DGH}_{DT} = \text{DGH}_{D_1} \times \ldots \times \text{DGH}_{D_n}$

Domain generalization hierarchy – Example

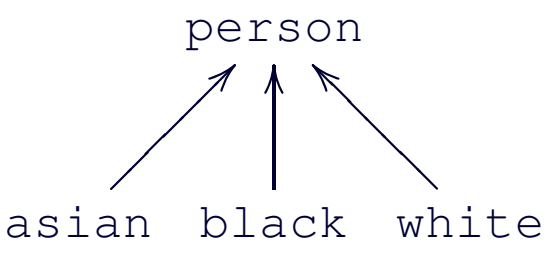




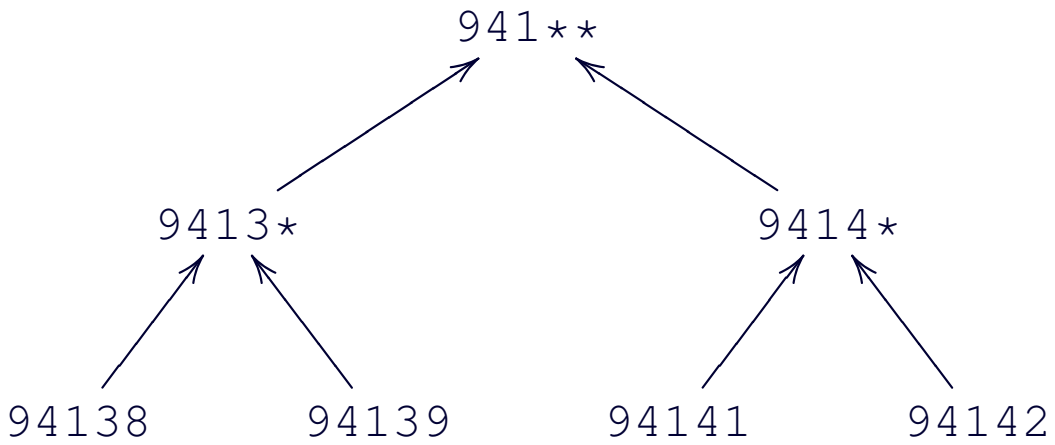
Value generalization hierarchy

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

Value generalization hierarchy – Example



VGH_{R_0}



VGH_{Z_0}

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

Generalized table with suppression – Example

Race	ZIP
asian	94142
asian	94141
asian	94139
asian	94139
asian	94139
black	94138
black	94139
white	94139
white	94141

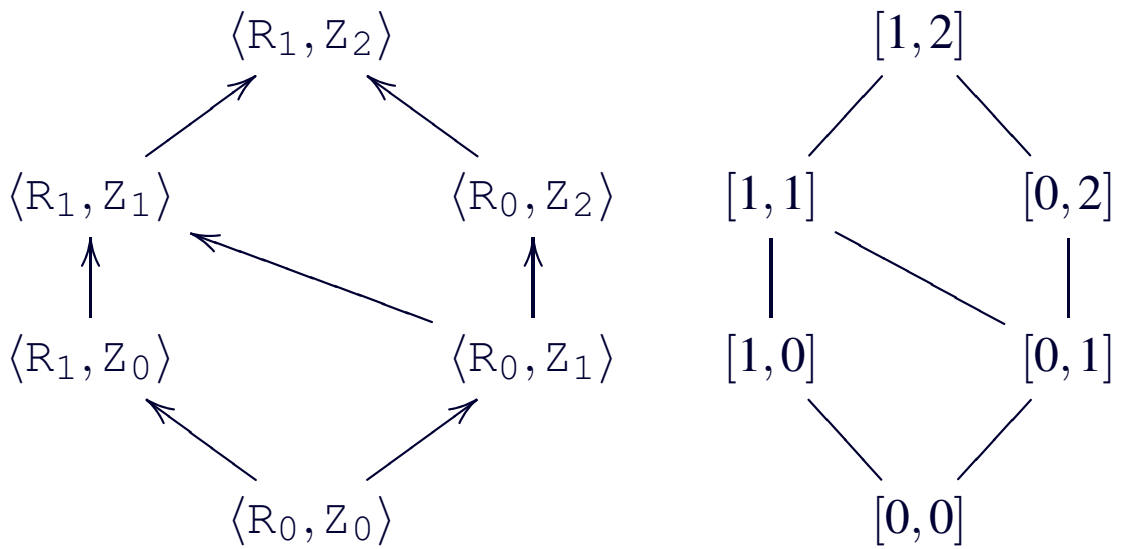
PT

Race	ZIP
person	94141
person	94139
person	94139
person	94139
person	94139
person	94139
person	94141

GT

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 $\mathrm{dom}(A_z,T_i)$ and $\mathrm{dom}(A_z,T_j)$ 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 $\Longrightarrow |T_j| \geq |T_z|$

2. $|T_i| - |T_j| \leq$ MaxSup

3. $\forall T_z : T_i \preceq T_z$ and T_z satisfies conditions 1 and 2 $\Longrightarrow \neg(DV_{i,z} < DV_{i,j})$

Examples of 2-minimal generalizations

MaxSup=2

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)

	Suppression			
Generalization	*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

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**

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	*

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_$\equiv$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**

CG_≡CG_CS

Race	DOB	Sex	ZIP

_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	*

_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

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 (e.g., [S-01], [BA-05], [LDR-05], [LDR-06])

k-anonymity revisited [GMT-08]

- 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 respondents
- When generalization is performed at attribute level (**AG**) this is equivalent to require each quasi-identifier n-uple to have at least k occurrences
- When generalization is performed at cell level (**CG**) the existence of at least k occurrences is a sufficient but not necessary condition; a less strict requirement would suffice
  1. for each sequence of values pt in PT[QI] there are at least k tuples in GT[QI] that contain a sequence of values generalizing pt
  2. for each sequence of values t in GT[QI] there are at least k tuples in PT[QI] that contain a sequence of values for which t is a generalization

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	94141
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 of the sensitive attribute values

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

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

Background knowledge

- Based on prior knowledge of some additional external information
  - an adversary knows that Hellen is a white female and she is in the microdata table
  - the adversary can infer that the disease of Hellen is either chest pain or short breath
  - the adversary knows that the Hellen runs 2 hours a day and therefore that Hellen cannot suffer from short breath
    $\Longrightarrow$ the adversary infers 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 (1)

- 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 "well-represented" values for the sensitive attribute in T
  - "well-represented" different definitions based on entropy or recursion (e.g., a q-block is ℓ-diverse if removing a sensitive value it remains (ℓ-1)-diverse)
- ℓ-diversity: an adversary needs to eliminate at least ℓ-1 possible values to infer that a respondent has a given value

ℓ-diversity (2)

- T is ℓ-diverse if all its q-blocks are ℓ-*diverse*

  $\Longrightarrow$ the homogeneity attack is not possible anymore
  $\Longrightarrow$ the background knowledge attack becomes more difficult

- 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 occurs when the distribution in a q-block is different from the distribution in the original population
- 20% of the population suffers from diabetes; 75% of tuples in a q-block have diabetes
  $\Longrightarrow$ people in the q-block have higher probability of suffering 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

Group 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, which however might be difficult to achieve

External Knowledge

External knowledge [CLR-07,MKMGH-07] (1)

- The consideration of the adversary's background knowledge (or external knowledge) is necessary when reasoning about privacy in data publishing
- External knowledge can be exploited for inferring sensitive information about individuals with high confidence
- Positive inference
  - a respondent has a given value (or a value within a restricted set)
- Negative inference
  - a respondent does not have a given value
- Existing approaches have mostly focused on positive inference

External knowledge (2)

- External knowledge may include:
  - similar datasets released by different organizations
  - instance-level information
  - ...
- Not possible to know a-priori what external knowledge the adversary possesses
- It is necessary to provide the data owner with a means to specify adversarial knowledge

External knowledge modeling [CLR-07]

- An adversary has knowledge about an individual (target) represented in a released table and knows the individual's QI values
  $\Longrightarrow$ goal: predict whether the target has a target sensitive value
- External knowledge modeled through a logical expression
- Three basic classes of expressions, representing knowledge about:
  - the target individual: information that the adversary may know about the target individual
  - others: information about individuals other than the target
  - same-value families: knowledge that a group (or family) of individuals have the same sensitive value
- Other types of external knowledge may be identified. . . . . .

External knowledge – Example (1)

Name	DOB	Sex	ZIP	Disease
Alice	74/04/12	F	94142	aids
Bob	74/04/13	M	94141	flu
Carol	74/09/15	F	94139	flu
David	74/03/13	M	94139	aids
Elen	64/03/18	F	94139	flu
Frank	64/09/27	M	94138	short breath
George	64/09/27	M	94139	flu
Harry	64/09/27	M	94139	aids

Original table

$\Longrightarrow$

DOB	Sex	ZIP	Disease
74	*	941**	aids
74	*	941**	flu
74	*	941**	flu
74	*	941**	aids
64	*	941**	flu
64	*	941**	short breath
64	*	941**	flu
64	*	941**	aids

4-anonymized table

Released table is 4-anonymized but ……

External knowledge – Example (2)

DOB	Sex	ZIP	Disease
74	*	941**	aids
74	*	941**	flu
74	*	941**	flu
74	*	941**	aids
64	*	941**	flu
64	*	941**	short breath
64	*	941**	flu
64	*	941**	aids

4-anonymized table

An adversary knows that Harry, born in 64 and living in area 94139, is in the table

External knowledge – Example (2)

DOB	Sex	ZIP	Disease
74	*	941**	aids
74	*	941**	flu
74	*	941**	flu
74	*	941**	aids
64	*	941**	flu
64	*	941**	short breath
64	*	941**	flu
64	*	941**	aids

4-anonymized table

$\Longrightarrow$

DOB	Sex	ZIP	Disease
64	*	941**	flu
64	*	941**	short breath
64	*	941**	flu
64	*	941**	aids

4-anonymized table

An adversary knows that Harry, born in 64 and living in area 94139, is in the table

$\Longrightarrow$ Harry belongs to the secong group

$\Longrightarrow$ Harry has aids with confidence 1/4

External knowledge – Example (3)

DOB	Sex	ZIP	Disease
64	*	941**	flu
64	*	941**	short breath
64	*	941**	flu
64	*	941**	aids

4-anonymized table

From another dataset, the adversary knows that George (who is in the table, is born in 64, and leaves in area 941**) has flu

External knowledge – Example (3)

DOB	Sex	ZIP	Disease
64	*	941**	flu
64	*	941**	short breath
64	*	941**	flu
64	*	941**	aids

4-anonymized table

$\Longrightarrow$

DOB	Sex	ZIP	Disease
64	*	941**	short breath
64	*	941**	flu
64	*	941**	aids

4-anonymized table

From another dataset, the adversary knows that George (who is in the table, is born in 64, and leaves in area 941**) has flu

$\Longrightarrow$ Harry has aids with confidence 1/3

External knowledge – Example (4)

DOB	Sex	ZIP	Disease
64	*	941**	short breath
64	*	941**	flu
64	*	941**	aids

4-anonymized table

From personal knowledge, the adversary knows that Harry does not have short breath

External knowledge – Example (4)

DOB	Sex	ZIP	Disease

$\Longrightarrow$

DOB	Sex	ZIP	Disease
64	*	941**	short breath
64	*	941**	flu
64	*	941**	aids

4-anonymized table

DOB	Sex	ZIP	Disease
64	*	941**	flu
64	*	941**	aids

4-anonymized table

From personal knowledge, the adversary knows that Harry does not have short breath

$\Longrightarrow$ Harry has aids with confidence 1/2

Multiple releases

- Data may be subject to frequent changes and may need to be published on regular basis

- The multiple release of a microdata table may cause information leakage since a malicious recipient can correlate the released datasets

Multiple independent releases – Example (1)

T_1

DOB	Sex	ZIP	Disease
74	*	941**	aids
74	*	941**	flu
74	*	941**	flu
74	*	941**	aids
64	*	941**	flu
64	*	941**	short breath
64	*	941**	flu
64	*	941**	aids

4-anonymized table at time t_1

T_2

DOB	Sex	ZIP	Disease
[70-80]	F	9414*	hypertension
[70-80]	F	9414*	gastritis
[70-80]	F	9414*	aids
[70-80]	F	9414*	gastritis
[60-70]	M	9413*	flu
[60-70]	M	9413*	aids
[60-70]	M	9413*	flu
[60-70]	M	9413*	gastritis

4-anonymized table at time t_2

An adversary knows that Alice, born in 1974 and living in area 94142, is in both releases

Multiple independent releases – Example (1)

T_1

DOB	Sex	ZIP	Disease
74	*	941**	aids
74	*	941**	flu
74	*	941**	flu
74	*	941**	aids

4-anonymized table at time t_1

T_2

DOB	Sex	ZIP	Disease
[70-80]	F	9414*	hypertension
[70-80]	F	9414*	gastritis
[70-80]	F	9414*	aids
[70-80]	F	9414*	gastritis

4-anonymized table at time t_2

An adversary knows that Alice, born in 1974 and living in area 94142, is in both releases

$\Longrightarrow$ Alice belongs to the first group in T_1

$\Longrightarrow$ Alice belongs to the first group in T_2

Multiple independent releases – Example (1)

T_1

DOB	Sex	ZIP	Disease
74	*	941**	aids
74	*	941**	flu
74	*	941**	flu
74	*	941**	aids

4-anonymized table at time t_1

T_2

DOB	Sex	ZIP	Disease
[70-80]	F	9414*	hypertension
[70-80]	F	9414*	gastritis
[70-80]	F	9414*	aids
[70-80]	F	9414*	gastritis

4-anonymized table at time t_2

An adversary knows that Alice, born in 1974 and living in area 94142, is in both releases

$\Longrightarrow$ Alice belongs to the first group in T_1

$\Longrightarrow$ Alice belongs to the first group in T_2

Alice suffers from aids (it is the only illness common to both groups)

Multiple independent releases – Example (2)

T_1

DOB	Sex	ZIP	Disease
74	*	941**	aids
74	*	941**	flu
74	*	941**	flu
74	*	941**	aids
64	*	941**	flu
64	*	941**	short breath
64	*	941**	flu
64	*	941**	aids

4-anonymized table at time t_1

T_2

DOB	Sex	ZIP	Disease
[70-80]	F	9414*	hypertension
[70-80]	F	9414*	gastritis
[70-80]	F	9414*	aids
[70-80]	F	9414*	gastritis
[60-70]	M	9413*	flu
[60-70]	M	9413*	aids
[60-70]	M	9413*	flu
[60-70]	M	9413*	gastritis

4-anonymized table at time t_2

An adversary knows that Frank, born in 1964 and living in area 94132, is in T_1 but not in T_2

Multiple independent releases – Example (2)

T_1

DOB	Sex	ZIP	Disease
64	*	941**	flu
64	*	941**	short breath
64	*	941**	flu
64	*	941**	aids

4-anonymized table at time t_1

T_2

DOB	Sex	ZIP	Disease
[60-70]	M	9413*	flu
[60-70]	M	9413*	aids
[60-70]	M	9413*	flu
[60-70]	M	9413*	gastritis

4-anonymized table at time t_2

An adversary knows that Frank, born in 1964 and living in area 94132, is in T_1 but not in T_2

Multiple independent releases – Example (2)

T_1

DOB	Sex	ZIP	Disease
64	*	941**	flu
64	*	941**	short breath
64	*	941**	flu
64	*	941**	aids

4-anonymized table at time t_1

T_2

DOB	Sex	ZIP	Disease
[60-70]	M	9413*	flu
[60-70]	M	9413*	aids
[60-70]	M	9413*	flu
[60-70]	M	9413*	gastritis

4-anonymized table at time t_2

An adversary knows that Frank, born in 1964 and living in area 94132, is in T_1 but not in T_2

$\Longrightarrow$ Frank suffers from short breath
(it is the only illness that appears in T_1 and does not appear in T_2)

m-invariance [XT-07]

A sequence $T_1, \ldots, T_n$ of released microdata tables satisfies m-invariance iff

- each equivalence class includes at least m tuples
- no sensitive value appears more than once in each equivalence class
- for each tuple t, the equivalence classes to which t belongs in the sequence are characterized by the same set of sensitive values

$\Longrightarrow$ the correlation of the tuples in $T_1, \ldots, T_n$ does not permit a malicious recipient to associate less than m different sensitive values with each respondent

k-anonymity in various applications

In addition to classical microdata release problem, the concept of k-anonymity and its extensions can be applied in different scenarios, e.g.:

- social networks (e.g.,[HMJTW-08])
- data mining (e.g.,[FWY-07, FWS-08])
- location data (e.g.,[GL-08])
- ...

Re-identification with any information

- Any information can be used to re-identify anonymous data

  $\Longrightarrow$ ensuring proper privacy protection is a difficult task since the amount and variety of data collected about individuals is increased

- Two examples:
  - AOL
  - Netflix

AOL data release (1)

- In 2006, to embrace the vision of an open research community, AOL (America OnLine) publicly posted to a website 20 million search queries for 650,000 users of AOL's search engine summarizing three months of activity
- AOL suppressed any obviously identifying information such as AOL username and IP address
- AOL replaced these identifiers with unique identification numbers (this made searches by the same user linkable)

AOL data release (2)

- User 44117749:
  - “numb fingers”, “60 single men”, “dog that urinates on everything”
  - “hand tremors”, “nicotine effects on the body”, “dry mouth”, and “bipolar”
  - “Arnold” (several people with this last name)
  - “landscapers in Lilburn, Ga”, “homes sold in shadow lake subdivision Gwinnett county, Georgia”

  $\Longrightarrow$ Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga

- She was re-identified by two New York Times reporters

- She explained in an interview that she has three dogs and that she searched for medical conditions of some friends

AOL data release (3)

What about user 17556639?

- how to kill your wife
- how to kill your wife
- wife killer
- how to kill a wife
- poop
- dead people
- pictures of dead people
- killed people
- dead pictures
- dead pictures
- dead pictures
- murder photo
- steak and cheese
- photo of death
- photo of death
- death
- dead people photos
- photo of dead people
- www.murderdpeople.com
- decapatated photos
- decapatated photos
- car crashes3
- car crashes3
- car crash photo

Netflix prize data study (1)

- In 2006, Netlix (the world largest online movie rental service), launched the "Netflix Prize" (a challenge that lasted almost three years)
  - Prize of us $ 1 million to be awarded to those who could provide a movie recommendation algorithm that beated Netflix's algorithm by 10%
- Netflix provided 100 million records revealing how nearly 500,000 of its users had rated movies from Oct.'98 to Dec.'05
- In each record Netflix disclosed the movie rated, the rating assigned (1 to 5), and the date of the rating

Netflix prize data study (2)

- Only a sample (one tenth) of the database was released
- Some ratings were perturbed (but not much to not alter statistics)
- Identifying information (e.g., usernames was removed), but a unique user identifier was assigned to preserve rating-to-rating continuity
- Release was not k-anonymous for any $k > 1$

Netflix prize data study (3)

- De-identified Netflix data can be re-identifed by linking with external sources (e.g., user ratings from IMDb users)
  - Knowing the precise ratings a person has assigned to six obscure (outside the top 500) movies, an adversary is able to uniquely identify that person 84% of the time
  - Knowing approximately when ($\pm$ 2 weeks) a person has rated six movies (whether or not obscure), an adversary is able to reidentify in 99% of the cases
  - Knowing two movies a user has rated, with precise ratings and rating dates ($\pm$ 3 days), an adversary is able to reidentify 68% of the users
- Movies may reveal your political orientation, religious views, or sexual orientations (Netflix was sued by a lesbian for breaching her privacy)

Differential Privacy

Differential privacy [D-06] (1)

Differential privacy has been proposed as an approach for protecting the privacy of individuals either represented or not represented in the released microdata table

- traditional solutions assume that privacy of individuals not included in the dataset is not at risk

Differential privacy (2)

- Differential privacy tries to capture the following definition of privacy:
  - anything that can be learned about a respondent from the statistical database should be learnable without access to the database
- Only an empty dataset can guarantee absolute protection against information leakage
- Differential privacy aims at preventing adversaries from being capable to detect the presence or absence of a given individual in a dataset
- Differential privacy defines a property on the data release mechanism

Differential privacy (3)

Informally:

- Differential privacy requires the probability distribution on the published results of an analysis to be "essentially the same" independent of whether an individual is represented or not in the dataset

- EXAMPLE: an insurance company consults a medical dataset to decide whether an individual is eligible for a contract
  $\Longrightarrow$ the presence of the tuple representing the individual does not affect the final decision

Formally:

- A randomized function K gives ε-differential privacy if for all data sets D and D' differing on at most one row, and all $S \subseteq \text{Range}(K)$, $\Pr[K(D) \in S] \leq \exp(\varepsilon) \times \Pr[K(D') \in S]$

Non-interactive and interactive scenarios

- Interactive scenario: evaluation of queries over a private dataset
  - ε-differential privacy is obtained adding random noise to query results
  - noise follows Laplace distribution $Lap(\Delta(f)/\varepsilon)$, with $\Delta(f)$ the maximum difference between the query result over D and over D'
- Non-interactive scenario: public release of a dataset
  - traditionally consists in the release of a frequency matrix
  - each cell is the result of a count query
  - ε-differential privacy is obtained adding random noise to each cell

$\Longrightarrow$ data truthfulness is not preserved

Relaxing differential privacy

- ε-differential privacy imposes a strict constraint

  $\Longrightarrow$ released data are noisy

- Relaxed privacy requirements permit the data recipient to benefit from more precise (i.e., with less additional noise) datasets
  - (ε, δ)-differential privacy [DS-09]: the ε bound on query answer probabilities may be violated with small probability (controlled by δ)
  - Computational differential privacy [MPRV-09]: a computationally bounded adversary should not be able to distinguish the query results computed over D from the ones computed over D'

Differential privacy for count queries

Differential privacy

- does not guarantee accuracy for queries involving a high number of respondents
  - variance of the additional noise for each cell $\Theta(1)$
  - variance of the additional noise for a query involving m cells $\Theta(m)$
- does not take into account correlated queries
  - two evaluations of the same query should provide the same result

$\Longrightarrow$ it is necessary to define specific approaches that overcome the above limitations (e.g., [LHRMM-10], [XWG-11])

Differential privacy in various applications

Similarly to k-anonymity, differentially private mechanisms have been developed for different domains, e.g.:

- social networks (e.g., [HLMJ-09, MW-09, RHMS-09])
- data mining (e.g., [CMFDX-11, DWHL-11, MCFY-11])
- location data (e.g., [HR-11])
- ...

Is differential privacy enough?

- Limiting the inference about the presence of a tuple is different from limiting the inference about the participation of the individual in the data generating process [KM-11, KM-12]
  - Bob's participation in a social network can cause links to form between Bob's friends (Bob's participation affects more than just the tuple marked "Bob")
- Differential privacy composes well with itself but not necessarily with other privacy definitions or data release mechanisms (which represent background knowledge that can cause privacy breaches)

Some open issues

- New privacy metrics
- New techniques to protect privacy
- External knowledge and adversarial attacks
- Evaluation of privacy vs utility

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 2005)*, 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.
- [AW-89] N.R. Adam, J.C. Wortmann, "Security-control methods for statistical databases: A comparative study," in *ACM Computing Survey*, vol. 21, n. 4, December 1989, pp. 515-556.
- [BA-05] R.J. Bayardo, R. Agrawal, "Data privacy through optimal k-anonymization," in *Proc. of the 21st International Conference on Data Engineering (ICDE 2005)*, Tokyo, Japan, 2005.
- [CMFDX-11] R. Chen, N. Mohammed, B.C.M. Fung, B.C. Desai, L. Xiong, "Publishing set-valued data via differential privacy," in *PVLDB,* 4(11):1087-1098, September 2011.
- [CDFS-07] 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.

References (2)

- [CLR-07] B-C. Chen, K. LeFevre, R. Ramakrishnan, "Privacy skyline: Privacy with multidimensional adversarial knowledge," in *Proc. of the 33rd International Conference on Very Large Data Bases (VLDB 2007)*, Vienna, Austria, 2007.
- [DWHL-11] B. Ding, M. Winslett, J. Han, Z. Li, "Differentially private data cubes: Optimizing noise sources and consistency," in *Proc. of SIGMOD 2011*, Athens, Greece, June 2011.
- [D-06] C. Dwork, "Differential privacy," in *Proc. of the 33rd International Colloquium on Automata, Languages and Programming (ICALP 2006)*, Venice, Italy, 2006.
- [DS-09] C. Dwork, A. Smith, "Differential Privacy for Statistics: What We Know and What We Want to Learn," in *Journal of Privacy and Confidentiality,* 1(2):135-154, 2009.
- [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 2005)*, Tokyo, Japan, 2005.
- [FZ-08] K.B. Frikken, Y. Zhang, "Yet another privacy metric for publishing micro-data," In *Proc. of the 7th Workshop on Privacy in Electronic Society (WPES 2008)*, Alexandria, VA, USA, 2008.

References (3)

- [FWY-07] B.C.M. Fung, K. Wang, P.S. Yu, "Anonymizing classification data for privacy preservation," in *IEEE TKDE,* 19(5):711-725, May 2007.
- [FWS-08] A. Friedman, R. Wolff, A. Schuster, "Providing k-anonymity in data mining," in *the VLDB Journal,* 17(4):789-804, July 2008.
- [GL-08] B. Gedik, L. Liu, "Protecting location privacy with personalized k-anonymity: Architecture and algorithms," in *IEEE TMC,* 7(1):1-18, January 2008.
- [GMT-08] A. Gionis, A. Mazza and T. Tassa, "k-Anonymization revisited," in *Proc. of the International Conference on Data Engineering,* Cancun, Mexico, 2008.
- [HMJTW-08] M. Hay, G. Miklau, D. Jensen, D. Towsley, P. Weis, "Resisting structural re-identification in anonymized social networks," in *PVLDB,* 1(1):102-114, August 2008.
- [KG-06] D. Kifer, J. Gehrke, "Injecting utility into anonymized datasets," in *ACM SIGMOD International Conference on Management of Data*, Chicago, IL, USA, 2006.
- [KM-11] D. Kifer, A. Machanavajjhala, "No free lunch in data privacy," in *Proc. of SIGMOD 2011,* Athens, Greece, June 2011.

References (4)

- [KM-12] D. Kifer, A. Machanavajjhala, “A rigorous and customizable framework for privacy,” in *Proc. of PODS 2012,* Scottsdale, AZ, USA, May 2012.
- [I-02] V. Iyengar, “Transforming data to satisfy privacy constraints,” in *Proc. of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Alberta, Canada, 2002.
- [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. 49-60, Baltimore, MA, USA, 2005.
- [LDR-06] K. LeFevre, D.J. DeWitt, R. Ramakrishnan, “Mondrian multidimensional k-anonymity,” in *Proc. of the International Conference on Data Engineering (ICDE 2006)*, Atlanta, GA, USA, 2006.
- [LHRMM-10] C. Li, M. Hay, V. Rastogi, G. Miklau, A. McGregor, “Optimizing linear counting queries under differential privacy,” in *Proc. of PODS 2010*, Indianapolis, IN, USA, June 2010.
- [LLV-07] N. Li, T. Li, and S. Venkatasubramanian. “t-closeness: Privacy beyond k-anonymity and ℓ-diversity,” In *Proc. of the IEEE 23nd International Conference on Data Engineering (ICDE 2007)*, Istanbul, Turkey, 2007.

References (5)

- [MGK-06] A. Machanavajjhala, J. Gehrke, D. Kifer "ℓ-diversity: Privacy beyond k-anonymity," in *Proc. of the International Conference on Data Engineering (ICDE 2006)*, Atlanta, GA, USA, 2006.
- [MKMGH-07] D.J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, J.Y. Halpern, "Worst-case background knowledge for privacy-preserving data publishing," in *Proc. of the IEEE 23rd International Conference on Data Engineering (ICDE 2007)*, Istanbul, Turkey, 2007.
- [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.
- [MPRV-09] I. Mironov, O. Pandey, O. Reingold, S.P. Vadhan, "Computational Differential Privacy," in *Proc. of CRYPTO 2009,* Santa Barbara, CA, USA, August 2009.
- [MCFY-11] N. Mohammed, R. Chen, B.C.M. Fung, P.S. Yu, "Differentially private data release for data mining," in *Proc. of KDD 2011,* San Diego, CA, USA, August 2011.
- [NAC-07] M.E. Nergiz, M. Atzori, C. Clifton, "Hiding the presence of individuals from shared databases," in *Proc. of the 2007 ACM SIGMOD International Conference on Management of Data (SIGMOD 2007)*, Beijing, China, 2007.

References (6)

- [NCA-07] M. Nergiz, C. Clifton, A. Nergiz, "Multirelational k-anonymity," in *Proc. of the 23rd IEEE International Conference on Data Engineering (ICDE 2007)*, Istanbul, Turkey, 2007.
- [PTLX-09] J. Pei, Y. Tao, J. Li, X. Xiao, "Privacy preserving publishing on multiple quasi-identifiers," in *Proc. of the 25th IEEE International Conference on Data Engineering (ICDE 2009)*, Shanghai, China, 2009.
- [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.
- [TMK-08] M. Terrovitis, N. Mamoulis, P. Kalnis, "Privacy-preserving anonymization of set-valued data," *Proc. of the VLDB Endowment*, vol. 1, August 2008, pp. 115-125.
- [WF-06] K. Wang, B. Fung, "Anonymizing sequential releases," in *Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006)*, Philadelphia, PA, USA, 2006.
- [WXWF-10] K. Wang, Y. Xu, R. Wong, A. Fu, "Anonymizing temporal data," in *Proc. of the 2010 IEEE International Conference on Data Mining (ICDM 2010)*, Sydney, Australia, 2010.

References (7)

- [XT-06] X. Xiao, Y. Tao, "Personalized privacy preservation," in *Proc. of the 2006 ACM SIGMOD International Conference on Management of Data (SIGMOD 2006)*, Chicago, IL, USA, 2006.
- [XT-07] X. Xiao, Y. Tao, "m-invariance: Towards privacy preserving re-publication of dynamic datasets," in *Proc. of the 2007 ACM SIGMOD International Conference on Management of Data (SIGMOD 2007)*, Beijing, China, 2007.
- [XWG-11] X. Xiao, G. Wang, J. Gehrke, "Differential privacy via wavelet transforms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, n. 8, August 2011, pp. 1200–1214.
- [ZHPJTJ-09] B. Zhou, Y. Han, J. Pei, B. Jiang, Y. Tao, Y. Jia, "Continuous privacy preserving publishing of data streams," in *Proc. of the 12th International Conference on Extending Database Technology (EDBT 2009)*, Saint Petersburg, Russia, 2009.