Data Privacy Management 2010

Towards Knowledge Intensive Data Privacy

Vicenç Torra[1]

September 23, 2010

[1] Institut d'Investigació en Intel·ligència Artificial (IIIA-CSIC)

Part of this work was done with G. Navarro-Arribas, I. Cano and D. Abril

Introduction

- Methods and tools for data privacy

Introduction

- Methods and tools for data privacy

  ... introducing knowledge in these tools

Introduction

- Methods and tools for data privacy

  ... introducing knowledge in these tools

⇒ extensive use of additional knowledge in

- disclosure risk assessment
- data protection

Introduction

- The role of knowledge in data privacy
  - Existing tools use only limited information
    - The file to be protected in data protection
    - Original and protected files in risk assessment

Introduction

- The role of knowledge in data privacy
  - Existing tools use only limited information
    - The file to be protected in data protection
    - Original and protected files in risk assessment

  - However, privacy depends on the context
    and about the available information to intruders
    ⇒ Need of adding the semantic context

Introduction

- The role of knowledge in data privacy
  - Knowledge in data protection
    - Explicit representation of knowledge related to the data
      - ◦ protected data should satisfy the data models
      - → negative ages
      - ◦ protected data should permit meanginful analysis
      - → random generalization of ZIP codes or cities
    - *Semantic deepth* of categorical data
      - ◦ protection methods should take into account the semantics

Introduction

- The role of knowledge in data privacy
  - Knowledge in risk assessment
    - Consideration of related databases (with different schemas)
    - Consideration of related information from e.g. the web
      - ◦ schema matching, database integration technologies, ontologies, ontology matching, ...
      - → otherwise unlinkable databases: no risk detected

Introduction

- The role of knowledge in data privacy:

  Summarization
  - Knowledge intensive data protection methods
    improves the quality of the protected data,
    extending their application domain and simplifying its use.

Introduction

Some particular examples

- Constrained data
- Semantic data protection
- Knowledge-rich disclosure risk assessment

Introduction

Some particular examples

- Constrained data
  - age is positive
  - total income is the sum of basic salary plus incentives

Introduction

Some particular examples

- Constrained data
  - age is positive
  - total income is the sum of basic salary plus incentives

    - protection procedures compliant with the constraints

Introduction

Some particular examples

- Semantic data protection
  - Terms in natural language have some semantic meaning.

Outline

Introduction

Some particular examples

- Semantic data protection
  - Terms in natural language have some semantic meaning.

◦ protection methods using ontologies

→ methods for k-anonymity using dendrograms of categories

→ microaggregation using ontologies

(e.g. Wordnet or open directory project)

Introduction

Some particular examples

- Knowledge-rich disclosure risk assessment
  - Record linkage is a versatile tool for measuring disclosure risk
    $\Rightarrow$ even applicable to synthetic data

Outline

Introduction

Some particular examples

- Knowledge-rich disclosure risk assessment
  - Record linkage is a versatile tool for measuring disclosure risk
    $\Rightarrow$ even applicable to synthetic data

◦ New approaches for record linkage in new scenarios:

$\longrightarrow$ Record linkage for intruder's file with a different scheme / data

$\longrightarrow$ Record linkage taking into account how data has been protected

$\longrightarrow$ Supervised record linkage, and parameter determination

Introduction

Some particular examples

- **Constrained data → microaggregation**
- Semantic data protection
- Knowledge-rich disclosure risk assessment

Constrained Data

Introduction

- Edit constraints
- ... and microaggregation

Introduction

- When data is edited, variables satisfy some constraints,
- Application of masking methods,
  ... causes the violation of the constraints

Introduction

- Is microaggregation appropriate ?
- Constrained microaggregation.
  $\longrightarrow$ suitability
  $\longrightarrow$ characterization (options) for microaggregation

Outline

Outline

Motivation

Data Privacy (I)

Data Privacy:

- Data is perturbated before publication
- Perturbation: minimal to maintain data utility (information loss)
- Perturbation: but enough to ensure data privacy

Measures:

- Information Loss or Data Utility Measures (IL)
  - The smaller the loss, the better
- Disclosure Risk Measures (DR)
  - The smaller the risk, the better

However:

- IL and DR are in contradiction ($\text{Score} = (\text{IL} + \text{DR})/2$)
- Good method, if a good score / trade-off

Data Privacy (II)

Methods for Data Privacy:

- Different methods have been proposed for data privacy
  - Perturbative methods
    - Data is modified adding some noise
  - Non-perturbative methods
    - Data is modified but no noise is included
      (e.g., change of granularity)
  - Synthetic data generators
    - Data is *artificial* (disclosure risk is not avoided)

Data Privacy (III)

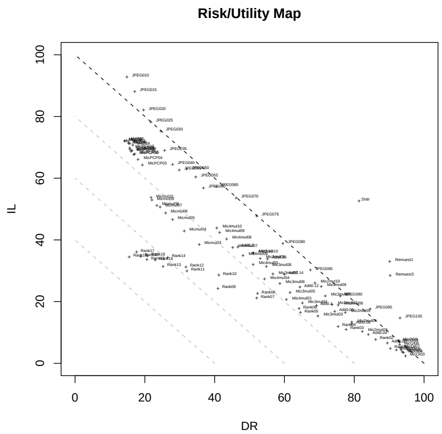
Methods for Data Privacy:

- Different methods have been proposed for data privacy
  - Perturbative methods
    - Microaggregation, rank swapping, ...
  - Non-perturbative methods
    - Suppression, top coding, most implementations for k-anonymity
  - Synthetic data generators
    - IPSO, Approach based on Fuzzy c-regression

Data Privacy (IV)

Methods for Data Privacy:

- Comparison of the methods:
  - U.S. Census Data Set: 1080 records, 13 variables
  - Score of around 30 (http://www.ppdm.cat):
  - Best performance: Microaggregation and rank swapping



Edit Constraints (I)

Constraints on the variables:

- **Linear constraints:**
  - E.g.,

$$\text{EC-LC1: } net + tax = gross$$

- Usual approach:
  - (i) edit data
  - (ii) protect data
  - (iii) edit again

    (to correct problems in protected data:

    some properties of the data protection method might be lost)

Edit Constraints (II)

Constraints on the variables:

- **Linear constraints:**
  - E.g.,

$$\text{EC-LC1: } net + tax = gross$$

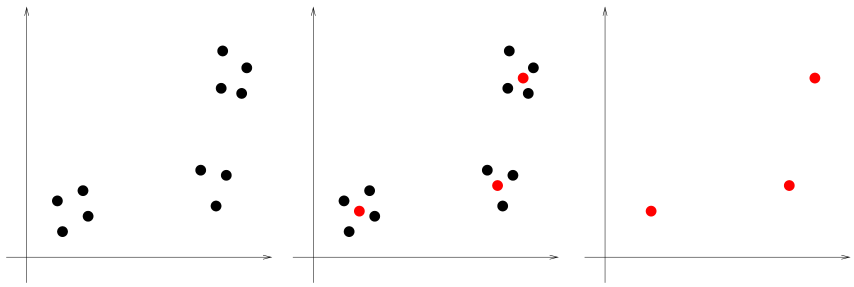
- **Is microaggregation appropriate ?**
  - Constrained microaggregation (to avoid new edition)
  - How aggregation should be done (in a sound way)?
  - Automate the process

Microaggregation

Microaggregation (I)

Microaggregation: Informal description

- (i) Build microclusters
- (ii) aggregate the records,
- (iii) replace records by aggregates
  - Privacy is ensured requiring k records in each cluster
  - Low information loss as clusters are small



Microaggregation (II)

Microaggregation: A formal description

- Notation.
  - $u_{ij} \in \{0, 1\}$ a partition: $u_{ij} = 1$ iff record j is assigned to the ith cluster.
  - v_i represents the ith cluster
  - k minimum number of records in a cluster, g number of clusters.
- Formalization.

Minimize $SSE = \sum_{i=1}^{g} \sum_{j=1}^{n} u_{ij}(d(x_j, v_i))^2$

Subject to $\sum_{i=1}^{g} u_{ij} = 1$ for all $j = 1, \ldots, n$

$2k \geq \sum_{j=1}^{n} u_{ij} \geq k$ for all $i = 1, \ldots, g$

$u_{ij} \in \{0, 1\}$

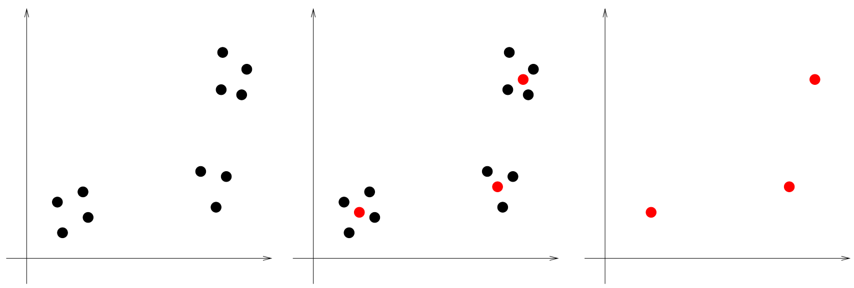
Microaggregation (III)

Microaggregation: Optimality

- Optimal NP hard for more than 2 variables
- Heuristic methods have been developed: MDAV

Microaggregation: Variations

- Fuzzy clustering-based Microaggregation:]
  - Avoids some adhoc attacks from intruders



Microaggregation (IV)

Microaggregation: The Operational approach.

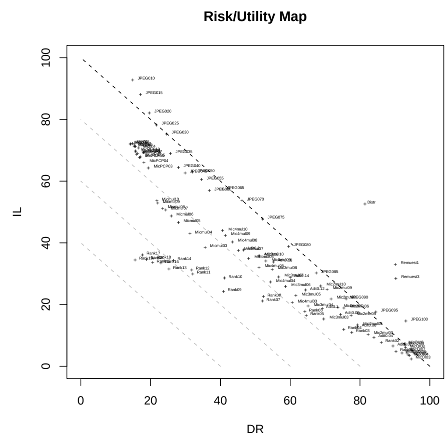
- 1. Clustering:
  Partition the set of records
  $\longrightarrow$ each partition element should have at least k records
- 2. Cluster representatives (aggregation):
  Compute a cluster representative for each cluster
- 3. Replacement:
  Replace each record by its cluster representative

Microaggregation (V)

Microaggregation: Discussion

- Microaggregation and k
  - The larger the k, the smaller the risk.
  - The larger the k, the larger the information loss.

    for a trade-off between risk and information loss $\Rightarrow$ find a good k



Microaggregation (VI)

Microaggregation: Discussion

- Microaggregation and k-anonymity
  - k-anonymity: k-indistinguishable records
  - Satisfied when all variables microaggregated together

    $\rightarrow$ microaggregation on the $\mathbb{R}^m$ space
  - Otherwise, in general, not satisfied.
  - Example

Microaggregation of data file	in terms of microaggregation of	and microaggregation of
(a_1, a_2, a_3, a_4)	(a_1, a_2)	(a_3, a_4)
(b_1, b_2, b_3, b_4)	(b_1, b_2)	(b_3, b_4)
(c_1, c_2, c_3, c_4)	(c_1, c_2)	(c_3, c_4)
...		
(z_1, z_2, z_3, z_4)	(z_1, z_2)	(z_3, z_4)

Edit Constraints

Introduction

- Edit constraints
  - A classification of the constraints

Edit Constraints

- Constraints on the possible values.
  - Values restricted to a predefined set
    - Values in a interval:

      EC-PV: $age \in [0, 125]$
  - Generalizable for subsets of variables
    - Values (v_1, v_2) in a subset of $D_1 \times D_2$

Edit Constraints

- One variable governs the possible values of another one
  - The values of a variable v_2 constrained by v_1
    - E.g., variable *sex* governing *number of pregnacies*

      EC-GV1: If *sex=male* THEN *number of pregnacies = 0*
    - or, e.g.[1]:

      EC-GV2: IF *age* < 17 THEN *gross income* $<$ *mean income*
    - or, e.g.[2]

      EC-GV3: *harvested acres* $\leq$ *planted acres*

[1] Shlomo, N., De Waal, T. (2008), Protection of micro-data subject to edit constraints against statistical disclosure, Journal of Official Statistics 24:2 229-253.

[2] Pierzchala, M. (1994) A review of the state of the art in automated data editing and imputation, in Statistical Data Editing, Vol. 1, Conference of European Statisticians Statistical Standards and Studies N. 44, United Nations Statistical Commission and Economic Commission for Europe, 10-40.

Edit Constraints

- Linear constraints.
  - Some variables satisfy some linear relationships.
    - E.g., *gross* in terms of *net* and *tax*

$$\text{EC-LC1: } \textit{net} + \textit{tax} = \textit{gross}$$

Edit Constraints

- Non-linear constraints.
  - The relationship between variables is not linear.
    - Relationship between *applicable VAT Rate*, *price exc. VAT*, and *retail price*:

      EC-NLC1:

      $$\textit{price exc. VAT} \cdot (\textit{1.00} + \textit{applicable VAT Rate}) = \textit{retail price}$$
    - Relationship between *wage sum*, *hours paid for*, and *wage rate*[3]:

      $$\text{EC-NLC2: } \textit{wage sum} = \textit{hours paid for} \cdot \textit{wage rate}$$

[3] Gasemyr, S. (2005) Editing and imputation for the creation of a linked micro file from base registers and other administrative data, Conference of European Statisticians, WP8.

Edit Constraints

- Other types of constraints.
  - E.g. constraints on categorical (ordinal or nominal) variables

Edit Constraints

- Values are restricted to exist in the domain
  - Values not only in the range but also exist in the data.
    - E.g. ages really existing in the population
      $\longrightarrow$ not enough to be in [0,125].
  - A perturbative method applied to data with ages in [0,30] should not lead to a file with a value equal to 50.
    - Application in linked files.

Microaggregation and Edit Constraints

Linear Constraints

Microaggregation and the edit constraints

- Microaggregation can deal easily with edit constraints
- Notation:
  - $x_1, \ldots, x_n$ records
  - $V_1, \ldots, V_m$ variables
  - $x_{i,j}$: value of record x_i for variable V_j

Microaggregation and the edit constraints

- Microaggregation and linear constraints:
  - Simplification on notation: V in terms of $V_1, \ldots, V_K$

V	V_1	$\cdots$	V_K
x_1	$x_{1,1}$	$\cdots$	$x_{1,K}$
$\vdots$	$\vdots$		$\vdots$
x_N	$x_{N,1}$	$\cdots$	$x_{N,K}$

  - Assumption$_1$: All the variables in the linear model are microaggregated together.
  - Assumption$_2$: Steps 1, 2, and 3 of the operational approach can be separated.
    $\longrightarrow$ cluster representative for each cluster satisfying the constraint

Microaggregation and the edit constraints

- Microaggregation and linear constraints:
  - Simplification on notation: V in terms of $V_1, \ldots, V_K$

V	V_1	$\cdots$	V_K
x_1	$x_{1,1}$	$\cdots$	$x_{1,K}$
$\vdots$	$\vdots$		$\vdots$
x_N	$x_{N,1}$	$\cdots$	$x_{N,K}$

  - Assumption$_3$: Linear constraint of the form $V = \sum_{i=1}^{K} \alpha_i V_i$
  - Naturally, the data also satisfies the constraints (i.e., the data were already edited). I.e.,
    $x_j = \sum_{i=1}^{K} \alpha_i x_{j,i}$ for all j.

Microaggregation and the edit constraints

- Microaggregation and linear constraints:
  - Simplification on notation: V in terms of $V_1, \ldots, V_K$

V	V_1	$\cdots$	V_K
x_1	$x_{1,1}$	$\cdots$	$x_{1,K}$
$\vdots$	$\vdots$		$\vdots$
x_N	$x_{N,1}$	$\cdots$	$x_{N,K}$
$\mathbb{C}(x_1, \ldots, x_N)$	$\mathbb{C}(x_{1,1}, \ldots, x_{N,1})$	$\cdots$	$\mathbb{C}(x_{1,K}, \ldots, x_{N,K})$

  - Assumption$_4$: The cluster representative is a function of the data in the cluster (each variable, independently): $\mathbb{C}$

Microaggregation and the edit constraints

- Microaggregation and linear constraints:
  - Simplification on notation: V in terms of $V_1, \ldots, V_K$

V	V_1	$\cdots$	V_K
x_1	$x_{1,1}$	$\cdots$	$x_{1,K}$
$\vdots$	$\vdots$		$\vdots$
x_N	$x_{N,1}$	$\cdots$	$x_{N,K}$
$\mathbb{C}(x_1, \ldots, x_N)$	$\mathbb{C}(x_{1,1}, \ldots, x_{N,1})$	$\cdots$	$\mathbb{C}(x_{1,K}, \ldots, x_{N,K})$

  - From these assumptions, we require:

$$\mathbb{C}(x_1, \ldots, x_N) = \sum_{i=1}^{K} \alpha_i \mathbb{C}(x_{1,i}, \ldots, x_{N,i})$$

Microaggregation and the edit constraints

- Microaggregation and linear constraints:
  - Simplification on notation: V in terms of $V_1, \ldots, V_K$

V	V_1	$\cdots$	V_K
x_1	$x_{1,1}$	$\cdots$	$x_{1,K}$
$\vdots$	$\vdots$		$\vdots$
x_N	$x_{N,1}$	$\cdots$	$x_{N,K}$
$\mathbb{C}(x_1, \ldots, x_N)$	$\mathbb{C}(x_{1,1}, \ldots, x_{N,1})$	$\cdots$	$\mathbb{C}(x_{1,K}, \ldots, x_{N,K})$

  - As $x_j = \sum_{i=1}^{N} \alpha_i x_{j,i}$ for all j in $\{1, \ldots, N\}$, we write:

$$\mathbb{C}(\sum_{i=1}^{K} \alpha_i x_{1,i}, \ldots, \sum_{i=1}^{K} \alpha_i x_{N,i}) = \sum_{i=1}^{K} \alpha_i \mathbb{C}(x_{1,i}, \ldots, x_{N,i})$$

Outline

Microaggregation and the edit constraints

- Microaggregation and linear constraints:
  - Simplification on notation: V in terms of $V_1, \ldots, V_K$

V	V_1	$\cdots$	V_K
x_1	$x_{1,1}$	$\cdots$	$x_{1,K}$
$\vdots$	$\vdots$		$\vdots$
x_N	$x_{N,1}$	$\cdots$	$x_{N,K}$
$\mathbb{C}(x_1, \ldots, x_N)$	$\mathbb{C}(x_{1,1}, \ldots, x_{N,1})$	$\cdots$	$\mathbb{C}(x_{1,K}, \ldots, x_{N,K})$

  - We also require reflexivity:

$$\mathbb{C}(x, \ldots, x) = x$$

Microaggregation and the edit constraints

- Microaggregation and linear constraints:
  - Proposition 1. (proof based on Functional Equations[4])
    $\mathbb{C}$ a function satisfying
    $$\mathbb{C}(\sum_{i=1}^{K} \alpha_i x_{1,i}, \ldots, \sum_{i=1}^{K} \alpha_i x_{N,i}) = \sum_{i=1}^{K} \alpha_i \mathbb{C}(x_{1,i}, \ldots, x_{N,i})$$
    for given values $\alpha_1, \ldots, \alpha_K$ ($\alpha_i \neq 0$) and arbitrary values $x_{i,j}$ for $1 \leq i \leq N$ and $1 \leq j \leq K$, and reflexivity
    $$\mathbb{C}(x, \ldots, x) = x$$
    Then, the most general solution for $\mathbb{C}$ is a function of the form
    $$\mathbb{C}(x_1, \ldots, x_N) = \sum_{i=1}^{N} \kappa_i x_i$$
    for κ_i such that $\sum_{i=1}^{N} \kappa_i = 1$ but otherwise arbitrary.

[4] Aczél, J. (1987) A Short Course on Functional Equations; J. Aczél (1966) Lectures on Functional Equations and their Applications, Academic Press.

Microaggregation and the edit constraints

- Microaggregation and linear constraints:
  - Proposition 2.
    $\mathbb{C}$ as before, but valid for all $\alpha_1, \ldots, \alpha_K$ ($\alpha_i \neq 0$):
    Same result:
    Then, the most general solution for $\mathbb{C}$ is a function of the form
    $$\mathbb{C}(x_1, \ldots, x_N) = \sum_{i=1}^{N} \kappa_i x_i$$
    for κ_i such that $\sum_{i=1}^{N} \kappa_i = 1$ but otherwise arbitrary.

Microaggregation and the edit constraints

- Microaggregation and linear constraints:
  - The only valid operator is a weighted mean
  - E.g., median is not valid for $V = V_1 + V_2$

V	V_1	V_2
3	1	2
6	0	6
8	2	6
6	1	6

Microaggregation and the edit constraints

- Microaggregation and linear constraints:
  - The only valid operator is a weighted mean
  - So the arithmetic mean is valid for $V = V_1 + V_2$ (i.e., WM with $\kappa_i = 1/3$)

V	V_1	V_2
3	1	2
6	0	6
8	2	6
17/3	3/3	14/3

Microaggregation and the edit constraints

- Microaggregation and linear constraints:
  - The number of elements in each partition element is not known
  - So, it is difficult to define *a priori* weights κ_i
  - In addition, the order of the elements should be irrelevant
- Proposition 3.
  - If we add symmetry:
    $$\mathbb{C}(x_1, \ldots, x_N) = \mathbb{C}(x_{\pi(1)}, \ldots, x_{\pi(N)})$$
    for an arbitrary permutation π, then the most general solution is
    $$\mathbb{C}(x_1, \ldots, x_N) = (1/N) \sum_{i=1}^{N} x_i$$

Microaggregation and the edit constraints

- Microaggregation and linear constraints:
  - The number of elements in each partition element is not known
  - So, it is difficult to define *a priori* weights κ_i
  - In addition, the order of the elements should be irrelevant
- An alternative: if $x_1 = x_2$, define $\kappa(x_1) = \kappa(x_2)$
  - According to Prop. 1, κ should be the same for all variables
  - The approach in most clustering algorithms follows this approach
  - E.g. in Fuzzy c-means for records $x_1, \ldots, x_N$ with memberships to the cluster equal to $\mu_1, \ldots, \mu_N$, $\rightarrow$ define
    $$\kappa_i = \frac{(\mu_i)^m}{\sum_{k=1}^{n} (\mu_k)^m}$$
    and then use the function $\mathbb{C}$.
  - This definition satisfies Prop. 1

Microaggregation and Edit Constraints

Nonlinear Constraints

Microaggregation and the edit constraints

- Microaggregation and nonlinear constraints:
  - We apply a similar approach:

V	V_1	$\ldots$	V_K
x_1	$x_{1,1}$	$\ldots$	$x_{1,K}$
$\vdots$	$\vdots$		$\vdots$
x_N	$x_{N,1}$	$\ldots$	$x_{N,K}$
$\mathbb{C}(x_1, \ldots, x_N)$	$\mathbb{C}(x_{1,1}, \ldots, x_{N,1})$	$\ldots$	$\mathbb{C}(x_{1,K}, \ldots, x_{N,K})$

  - Now,

$$\mathbb{C}(x_1, \ldots, x_N) = \prod_{i=1}^{K} \mathbb{C}(x_{1,i}, \ldots, x_{N,i})^{\alpha_i}$$

  - If the original data satisfy this constraint (i.e., $x_j = \prod_{i=1}^{N} x_{j,i}^{\alpha_i}$),

$$\mathbb{C}\left(\prod_{i=1}^{K} x_{1,i}^{\alpha_i}, \ldots, \prod_{i=1}^{K} x_{N,i}^{\alpha_i}\right) = \prod_{i=1}^{K} \mathbb{C}(x_{1,i}, \ldots, x_{N,i})^{\alpha_i}$$

Microaggregation and the edit constraints

- Microaggregation and nonlinear constraints:
  - Proposition 4.
    $\mathbb{C}$ a function satisfying
    $$\mathbb{C}(\prod_{i=1}^K x_{1,i}^{\alpha_i}, \ldots, \prod_{i=1}^K x_{N,i}^{\alpha_i}) = \prod_{i=1}^K \mathbb{C}(x_{1,i}, \ldots, x_{N,i})^{\alpha_i}$$
    for given values $\alpha_1, \ldots, \alpha_K$ ($\alpha_i \neq 0$) and arbitrary values $x_{i,j}$ for $1 \leq i \leq N$ and $1 \leq j \leq K$, and reflexivity
    $$\mathbb{C}(x, \ldots, x) = x$$
    Then, the most general solution for $\mathbb{C}$ is a function of the form
    $$\mathbb{C}(x_1, \ldots, x_N) = \prod_{i=1}^N x_i^{\kappa_i}$$
    for κ_i such that $\sum_{i=1}^N \kappa_i = 1$ but otherwise arbitrary.

Microaggregation and the edit constraints

- Microaggregation and nonlinear constraints:
  - Results similar to the linear case (Propositions 5 and 6):
    - Same function $\mathbb{C}$ when arbitrary $\alpha_1, \ldots, \alpha_K$
    - Equal weights when symmetry is added:
      $$\mathbb{C}(x_1, \ldots, x_N) = \prod_{i=1}^N x_i^{1/N}$$

Microaggregation and Edit Constraints

Constraints on the Values

Microaggregation and the edit constraints

- Linear constraints, and constraints on the values
  - Simple formulation: data define an interval
    - Cluster representative in the interval defined between the minimum and the maximum of the elements in the cluster (internality).

$$\min x_i \leq \mathbb{C}(x_1, \ldots, x_N) \leq \max_i$$

  - Proposition 7. Adding internality to Proposition 1:

$$\mathbb{C}(x_1, \ldots, x_N) = \sum_{i=1}^{N} \kappa_i x_i$$

    for κ_i such that $\sum_{i=1}^{N} \kappa_i = 1$ and $\kappa_i \geq 0$ but otherwise arbitrary.

Microaggregation and the edit constraints

- Nonlinear constraints, and constraints on the values
  - Simple formulation: data define an interval
    - Cluster representative in the interval defined between the minimum and the maximum of the elements in the cluster (internality).
      $$\min x_i \leq \mathbb{C}(x_1, \ldots, x_N) \leq \max_i$$
  - Proposition 8. Adding internality to Proposition 4:
    $$\mathbb{C}(x_1, \ldots, x_N) = \prod_{i=1}^{N} x_i^{\kappa_i}$$
    for κ_i such that $\sum_{i=1}^{N} \kappa_i = 1$ and $\kappa_i \geq 0$ but otherwise arbitrary.

Microaggregation and Edit Constraints

One variable governs the possible values of another variable

Microaggregation and the edit constraints

- One variable governs another one
  - We cannot constraint microaggregation so easily in this case.
  - Study in a case by case basis.
  - Examples (from 1st section):
    EC-GV1: If *sex=male* THEN *number of pregnacies = 0*
    EC-GV2: IF *age* < 17 THEN *gross income* $<$ *mean income*
    EC-GV3: *harvested acres* $\leq$ *planted acres*

Microaggregation and the edit constraints

- One variable governs another one
  - Study in a case by case basis: Case EC-GV3
    EC-GV3: *harvested acres* $\leq$ *planted acres*
  - General case for variables V_1 and V_2 ($V_1 \leq V_2$):

V_1	V_2	$\cdots$	V_K
$x_{1,1}$	$x_{1,2}$	$\cdots$	$x_{1,K}$
$\vdots$	$\vdots$		$\vdots$
$x_{N,1}$	$x_{N,2}$	$\cdots$	$x_{N,K}$
$\mathbb{C}(x_{1,1}, \ldots, x_{N,1})$	$\mathbb{C}(x_{1,2}, \ldots, x_{N,2})$	$\cdots$	$\mathbb{C}(x_{1,K}, \ldots, x_{N,K})$

  - Assumptions and results ...

Microaggregation and the edit constraints

- One variable governs another one
  - General case for variables V_1 and V_2 ($V_1 \leq V_2$):

V_1	V_2	$\cdots$	V_K
$x_{1,1}$	$x_{1,2}$	$\cdots$	$x_{1,K}$
$\vdots$	$\vdots$		$\vdots$
$x_{N,1}$	$x_{N,2}$	$\cdots$	$x_{N,K}$
$\mathbb{C}(x_{1,1},\ldots,x_{N,1})$	$\mathbb{C}(x_{1,2},\ldots,x_{N,2})$	$\cdots$	$\mathbb{C}(x_{1,K},\ldots,x_{N,K})$

  - a) We assume that V_1 and V_2 are microaggregated together.
  - b) If data has already been edited,
    $x_{i,1} \leq x_{i,2}$ for all records i
  - c) So, the condition can be formalized as:
    if $x_{i,1} \leq x_{i,2}$ for all records i, then
    $$\mathbb{C}(x_{1,1},\ldots,x_{N,1}) \leq \mathbb{C}(x_{1,2},\ldots,x_{N,2})$$
    That is, $\mathbb{C}$ is monotonic.

Microaggregation and the edit constraints

- One variable governs another one. Results:
  - a) We assume that V_1 and V_2 are microaggregated together.
  - b) If data has already been edited,
    $x_{i,1} \leq x_{i,2}$ for all records i
  - c) So, the condition can be formalized as:
    if $x_{i,1} \leq x_{i,2}$ for all records i, then
    $$\mathbb{C}(x_{1,1},\ldots,x_{N,1}) \leq \mathbb{C}(x_{1,2},\ldots,x_{N,2})$$
    That is, $\mathbb{C}$ is monotonic.
- $\mathbb{C}$ in Prop. 3, 6, 7, 8 are monotonic. So, appropriate here.
- Proposition (solutions) (and the particular cases: $\kappa_i = 1/N$):
  - $\mathbb{C}(x_1,\ldots,x_N) = \sum_{i=1}^N \kappa_i x_i$
  - $\mathbb{C}(x_1,\ldots,x_N) = \prod_{i=1}^N x_i^{\kappa_i}$
    for κ_i such that $\sum_{i=1}^N \kappa_i = 1$ and $\kappa_i \geq 0$

Microaggregation and the edit constraints

- One variable governs another one
  - Study in a case by case basis: Case EC-GV1 and EC-GV2
    EC-GV1: If *sex=male* THEN *number of pregnacies = 0*
    EC-GV2: IF *age* < 17 THEN *gross income* $<$ *mean income*
  - Partition the file (horizontally) and microaggregate each subset[5].
    EC-GV1: Partition $X = \{\Pi_1, \Pi_2\}$,
    Π_1 with *sex=male* and Π_2 with *sex=female*.
    $\rightarrow$ any function $\mathbb{C}$ s.t. $\mathbb{C}(0, \ldots, 0) = 0$ is appropriate
    EC-GV2: Partition $X = \{\Pi_1, \Pi_2\}$,
    Π_1 with *age* < 17 and Π_2 with *age* ≥ 17.
    $\rightarrow$ any monotonic function $\mathbb{C}$ is appropriate

[5] Similar to: Shlomo, N., De Waal, T. (2008), Protection of micro-data subject to edit constraints against statistical disclosure, Journal of Official Statistics 24:2 229-253.

Microaggregation and Edit Constraints

Values are restricted to exist in the domain

Microaggregation and the edit constraints

- Values are restricted to exist in the domain
  - In previous propositions, only possible when $\kappa_i = 1$ for a particular i.
  - In general, adding this constraint to previous propositions results into: a overconstrained problem $\rightarrow$ i.e., no solution exists
  - Considering this constraint but not the other, any order statistic as e.g. the median[6], or boolean max-min functions.

[6] as used in: Sande, G. (2002) Exact and approximate methods for data directed microaggregation in one or more dimensions, Int. J. of Unc., Fuzz. and Knowledge Based Systems 10:5 459-476.

Implementation and Example

Implementation and Example (I)

- Specification of XML edit constraints as Schematron rules
  - Data and rules in XML format are validated
  - Rules are parsed to identify the type of edit constraint
  - Microdata is processed accordingly
    - ∗ Variables involved in an edit constraint are grouped together
    - ∗ Appropriate microaggregate is then used

Implementation and Example (II)

- Example:
  - Census Data set: 1080 records, 13 numerical variables
  - Scenario 1: constraints are considered
  - Scenario 2: constraints are ignored

Implementation and Example (III)

Scenario 1				Scenario 2			
k	PIL	DR	SCORE	k	PIL	DR	SCORE
2	30.305	51.128	40.716	2	34.418	32.986	33.702
3	36.251	42.374	39.312	3	41.462	26.293	33.878
4	40.004	36.897	38.450	4	46.678	22.600	34.639
5	42.188	33.360	37.774	5	49.145	20.024	34.584
9	48.379	27.024	37.702	9	55.568	14.843	35.206
10	48.484	25.962	37.223	10	56.375	14.046	35.210
15	52.485	22.620	37.553	15	58.735	11.660	35.197
20	54.542	20.493	37.517	20	60.383	10.265	35.324
25	56.523	18.643	37.583	25	61.655	8.764	35.210
30	58.164	16.866	37.515	30	62.753	7.886	35.320
35	59.621	15.233	37.427	35	63.656	7.506	35.581
40	59.870	14.364	37.117	40	64.436	6.640	35.538
45	61.251	13.642	37.446	45	65.368	6.570	35.783
70	67.038	10.125	38.581	70	67.453	4.967	36.210

Conclusions

Conclusions

- Microaggregation is specially suited when constraints are considered
- Analysis of the approaches when defining the centroids

Advertisement (SPAM)

Transactions on Data Privacy

- Transactions on Data Privacy
  - Launched in 2008
  - Three issues per year
  - Indexed in ACM Digital Library, DBLP, MathSciNet, DOAJ
  - Support by Catalan Assoc. on Artificial Intelligence (ECCAI member) and the Unesco Chair on Data Privacy
  - Open Access (no charges for publication)
  - http://www.tdp.cat