Data-driven: Tabular data protection

Vicenç Torra

November, 2022

Department of Computing Sciences, Umeå University

Outline

Protection for tabular data

Tabular data

- Aggregates of data with respect to a few variables.
  - Aggregates of data can lead to disclosure

Data-driven protection methods

Data protection methods for aggregates / tabular data

- **Example.** File A with records about profession, town and salaries.
  - $P_1, M_1, 160$
  - $P_1, M_1, 200$
  - $P_2, M_1, 40$
  - $P_2, M_1, 60$
  - $P_2, M_1, 75$

Data-driven protection methods

Data protection methods for aggregates / tabular data

- **Example.** File A with records about profession, town and salaries.
  - $P_1, M_1, 160$
  - $P_1, M_1, 200$
  - $P_2, M_1, 40$
  - $P_2, M_1, 60$
  - $P_2, M_1, 75$
- From A, we build a two-dimensional aggregated table with frequencies, another with salaries.
  - $T_f : P \times M \to \mathbb{N}$
  - $T_s : P \times M \to \mathbb{N}$
  - Tables have subtotals, and totals.

Data-driven protection methods

Data protection methods for aggregates / tabular data

- **Example.** Aggregate table of frequencies. Ex. (Castro, 2012)

	P_1	P_2	P_3	P_4	P_5	Total
M_1	2	15	30	20	10	77
M_2	72	20	1	30	10	133
M_3	38	38	15	40	5	136
TOTAL	112	73	46	90	25	346

Data-driven protection methods

Data protection methods for aggregates / tabular data

- **Example.** Aggregate table of magnitudes

	P_1	P_2	P_3	P_4	P_5	Total
M_1	360	450	720	400	360	2290
M_2	1440	540	22	570	320	2892
M_3	722	1178	375	800	363	3438
TOTAL	2522	2168	1117	1770	1043	8620

Data-driven protection methods

Data protection methods for aggregates / tabular data

- Disclosure risk ?
  - Aggregated data does not avoid disclosure

Data-driven protection methods

Data protection methods for aggregates / tabular data

- Disclosure problems
  - **External attack.** Combining the informations of the two tables the adversary is able to infer some sensitive information. The single person working as P_3 in town M_2 has a salary of 22.

Data-driven protection methods

Data protection methods for aggregates / tabular data

- Disclosure problems
  - **Internal attack.** A person whose data is in the database is able to use the information of the tables to infer some sensitive information about other individuals. A person with work P_1 and living in M_1 attacks the data using his own salary. For example, if there are only two doctors in a town, each one will be able to find the salary of the other.

Data-driven protection methods

Data protection methods for aggregates / tabular data

- Disclosure problems
  - **Internal attack with dominance.** Internal attack of a large contributor. E.g., we have 5 people (P_5, M_3). If one of them has a salary of 350, then it is clear that the salary of the other four is at most 13.

Data-driven protection methods

Data protection methods for aggregates / tabular data

- On the disclosure and the public information
  - The frequency of a cell may be obtained from public information. No need of being it published. E.g., there is a single doctor in a town, only a few teachers, or companies in a sector (for finacial data).

Tabular data

- Privacy model / disclosure risk measure
- Data protection mechanism
- Information loss

Tabular data: privacy model

- **Rule (n,k)-dominance.** A cell is sensitive when n contributions represent more than the k fraction of the total. That is, the cell is sentitive when

$$\frac{\sum_{i=1}^{n} c_{\sigma(i)}}{\sum_{i=1}^{t} c_i} > k$$

where $\{\sigma(1),...,\sigma(t)\}$ is a permutation of $\{1,...,t\}$ such that $c_{\sigma(i-1)} \geq c_{\sigma(i)}$ for all $i = \{2,...,t\}$ (i.e., $c_{\sigma(i)}$ is the ith largest element in the collection $c_1,...,c_t$).

This rule is used with $n = 1$ or $n = 2$ and $k > 0.6$.

Tabular data: privacy model

- **Example.** $(1, 50)$ means that there is one individual that contributes to more than 50% of the content. In cell (P_5, M_3), 5 people and one of them has a salary of 350, over 363. As $350/363 > 0.5$: Cell is sensitive

	P_1	P_2	P_3	P_4	P_5	Total
M_1	360	450	720	400	360	2290
M_2	1440	540	22	570	320	2892
M_3	722	1178	375	800	363	3438
TOTAL	2522	2168	1117	1770	1043	8620

Tabular data: privacy model

- **Rule pq.** This rule is also known as the prior/posterior rule. It is based on two positive parameters p and q with $p < q$. Prior to the publication of the table, any intruder can estimate the contribution of contributors within the q percent. Then, a cell is considered sensitive if an intruder on the light of the released table can estimate the contribution of a contributor within p percent.

- **Rule p%.** This rule can be seen as a special case of the previous rule when no prior knowledge is assumed on any cell. Because of that, it can be seen as equivalent to the previous rule with $q = 100$.

Tabular data: privacy model

- **Rule pq.** This rule is also known as the prior/posterior rule. It is based on two positive parameters p and q with $p < q$. Prior to the publication of the table, any intruder can estimate the contribution of contributors within the q percent. Then, a cell is considered sensitive if an intruder on the light of the released table can estimate the contribution of a contributor within p percent.

- **Rule p%.** This rule can be seen as a special case of the previous rule when no prior knowledge is assumed on any cell. Because of that, it can be seen as equivalent to the previous rule with $q = 100$.

- **Example.** if contributions to (P_5, M_3) are $350, 4, 3, 3, 3$, this as $363 - 350 - 4 < 0.5 \cdot 363$, the rule implies: it is sensitive.

Tabular data: data protection mechanism

- Protection of a tabular data
  - Perturbative. values are modified
    - ⋆ Post-tabular. Noise added after table preparation
      - Rounding
      - Controlled tabular adjustment (CTA). Replacing a table by another that is *similar*
    - ⋆ Pre-tabular. Noise added before table preparation
  - Non-perturbative. cell suppression

Tabular data: data protection mechanism

- Protection of a tabular data: cell suppression
- Primary suppression not enough:

	P_1	P_2	P_3	P_4	P_5	Total
M_1	360	450	720	400	360	2290
M_2	1440	540	~~22~~	570	320	2892
M_3	722	1178	375	800	363	3438
TOTAL	2522	2168	1117	1770	1043	8620

- Secondary suppressions required: (Primary suppr. (PS), Secondary suppr. (SS)

	P_1	P_2	P_3	P_4	P_5	Total
M_1	360	450	SS	400	SS	2290
M_2	1440	540	PS	570	SS	2892
M_3	722	1178	375	800	363	3438
TOTAL	2522	2168	1117	1770	1043	8620

- Solutions built using optimization

Tabular data: data protection mechanism

- Protection of a tabular data: cell suppression
  - Decide which cells to suppress
  - Given a set of sensitive cells
  - Estimated values for suppressed cells should be outside a given interval
    (upper and lower protection levels;
    estimation based on non suppressed values + linear relationships)

  $\Rightarrow$ Problem formulated as an optimization problem

Tabular data: data protection mechanism

- Protection of a tabular data: cell suppression

$$\min \quad \sum_{i=1}^{n} w_i y_i$$

subject to

$$Ad^l = 0$$
$$(klo_i - a_i)y_i \leq d^{l,i} \leq (kup_i - a_i)y_i \quad \text{for all } i = 1, \ldots, n$$
$$d^{l,p} \leq -lo_p \quad \text{for all } p \in \mathcal{P}$$

$$Ad^u = 0$$
$$(klo_i - a_i)y_i \leq d^{u,i} \leq (kup_i - a_i)y_i \quad \text{for all } i = 1, \ldots, n$$
$$d^{u,p} \geq up_p \quad \text{for all } p \in \mathcal{P}$$

$$y_i \in \{0, 1\} \quad \text{for } i = 1, \ldots, n$$

Tabular data: information loss

- Minimal number of suppressions
- Weights associated to cells: *minimal weight* of suppressed cells

Summary

Summary

How to protect tabular data?

- Privacy models for tabular data
- Data protection mechanisms for tabular data:
  cell suppression
  controlled tabular adjustment