Data protection procedures

Vicenç Torra

January, 2018

SAIL + PICS, School of Informatics, University of Skövde, Sweden

Outline

Computation-driven approaches

Data Privacy

Respondent and owner privacy

- Data-driven or general-purpose
- Computation-driven or specific-purpose (Ch. 3.4)
  - Single database: differential privacy (Ch. 3.4.1)
  - Multiple databases:
    - ⋆ Centralized approach: trusted third party (Ch. 3.4.2)
    - ⋆ Distributed approach: multiparty computation (Ch. 3.4.2)
- Result-driven

Data Privacy

Computation-driven approaches/multiple databases: centralized

- **Example.** Parties $P_1, \ldots, P_n$ own databases $DB_1, \ldots, DB_n$. The parties want to compute a function, say f, of these databases (i.e., $f(DB_1, \ldots, DB_n)$) without revealing unnecessary information. In other words, after computing $f(DB_1, \ldots, DB_n)$ and delivering this result to all P_i, what P_i knows is nothing more than what can be deduced from his DB_i and the function f.
- So, the computation of f has not given P_i any extra knowledge.

Data Privacy

Computation-driven approaches/multiple databases: distributed

- Vertically partitioned data
- Horizontally partitioned data

Data Privacy

Computation-driven approaches/multiple databases: distributed

- The centralized approach as a reference

Data Privacy

Computation-driven approaches/multiple databases: distributed

Privacy leakage for the distributed approach is usually analyzed considering two types of adversaries.

Data Privacy

Computation-driven approaches/multiple databases: distributed

Privacy leakage for the distributed approach is usually analyzed considering two types of adversaries.

- **Semi-honest adversaries.** Data owners follow the cryptographic protocol but they analyse all the information they get during its execution to discover as much information as they can.
- **Malicious adversaries.** Data owners try to fool the protocol (e.g. aborting it or sending incorrect messages on purpose) so that they can infer confidential information.

Result-driven approaches

Data Privacy

Respondent and owner privacy

- Data-driven or general-purpose
- Computation-driven or specific-purpose
- Result-driven (Ch. 3.5)

Data Privacy

Result-driven

- Prevent data mining procedures infer some knowledge that is valuable for the database owner
- Other uses: avoid discriminatory knowledge inferred from databases

Data Privacy

Result-driven

- **Formalization.** Database $\mathcal{D}$, A data mining algorithm, with parameters Θ is said to have ability to derive knowledge K from $\mathcal{D}$ if and only if K is obtained from the output of the algorithm. Notation: $(A, \mathcal{D}, \Theta) \vdash K$.
- Any knowledge K such that $(A, \mathcal{D}, \Theta) \vdash K$ is in $KSet_{\mathcal{D}}$.

Data Privacy

Result-driven

- **Formalization.** Database $\mathcal{D}$, A data mining algorithm, with parameters Θ is said to have ability to derive knowledge K from $\mathcal{D}$ if and only if K is obtained from the output of the algorithm. Notation: $(A, \mathcal{D}, \Theta) \vdash K$.
- Any knowledge K such that $(A, \mathcal{D}, \Theta) \vdash K$ is in $KSet_{\mathcal{D}}$.

Definition. $\mathcal{D}$ a database, $\mathcal{K} = \{K_1, \ldots, K_n\}$ sensitive knowledge to be hidden. The problem of hiding knowledge $\mathcal{K}$ from $\mathcal{D}$ consists on transforming $\mathcal{D}$ into a database $\mathcal{D}'$ such that

1. $\mathcal{K} \cap KSet_{\mathcal{D}'} = \emptyset$
2. the information loss from $\mathcal{D}$ to $\mathcal{D}'$ is minimal

Data Privacy

Result-driven for association rules mining: Association rule hiding

- Recall that rules are mined when

$$Support(R) \geq thr - s$$

and

$$Confidence(R) \geq thr - c$$

for certain thresholds $thr - s$ and $thr - c$.

Data Privacy

Result-driven for association rules mining: Association rule hiding

- Recall that rules are mined when

$$Support(R) \geq thr - s$$

and

$$Confidence(R) \geq thr - c$$

for certain thresholds $thr - s$ and $thr - c$.

Two approaches:

- To reduce the support of the rule.
- To reduce the confidence of the rule.

Data Privacy

Result-driven for association rules mining: example

- **A formalization.** $\mathcal{D}$ a database; $thr - s$ threshold. Let $\mathcal{K} = \{K_1, \ldots, K_n\}$ sensitive itemsets, $\mathcal{A}$ non-sensitive itemsets.

Data Privacy

Result-driven for association rules mining: example

- **A formalization.** $\mathcal{D}$ a database; $thr - s$ threshold. Let $\mathcal{K} = \{K_1, \ldots, K_n\}$ sensitive itemsets, $\mathcal{A}$ non-sensitive itemsets.
- Transform $\mathcal{D} \rightarrow \mathcal{D}'$ such that
  1. $Support_{\mathcal{D}'}(K) < thr - s$ for all $K_i \in \mathcal{K}$
  2. The number of itemsets K in $\mathcal{A}$ such that $Support_{\mathcal{D}'}(K) < thr - s$ is minimized.

This problem is NP-hard (Atallah et al., 1999)

Because of this: heuristic approaches

Data Privacy

Result-driven for association rules mining: heuristic algorithm

- **Algorithm.**

  While HI is not hidden **do**
  HI' = HI;
  While $|HI'| > 2$ **do**
  P = subsets of HI with cardinality $|HI'| - 1$;
  HI'= $\arg\max_{hi \in P} Support(hi)$;
  Ts = transaction in T supporting HI that affects
  the mininum number of itemsets of cardinality 2;
  Set HI' = 0 in Ts;
  Propagate results forward;

Data Privacy

Result-driven for association rules mining: heuristic algorithm

- **Algorithm.**

  While HI is not hidden **do**
  HI' = HI;
  While $|HI'| > 2$ **do**
  P = subsets of HI with cardinality $|HI'| - 1$;
  HI'= $\arg\max_{hi \in P} Support(hi)$;
  Ts = transaction in T supporting HI that affects the mininum number of itemsets of cardinality 2;
  Set HI' = 0 in Ts;
  Propagate results forward;

- The algorithm does not cause false positives,
- only false negatives (rules no longer inferred)

Data Privacy

Result-driven for association rules mining: heuristic algorithm

- **Example.** Computation of the algorithm to hide $HI = \{a, b, c\}$.

Data Privacy

Result-driven for association rules mining: heuristic algorithm

- **Example.** Computation of the algorithm to hide $HI = \{a, b, c\}$.

Transaction number	Items
T1	a, b, c, d
T2	a, b, c
T3	a, c, d

Data Privacy

Result-driven for association rules mining: heuristic algorithm

- **Example.** Computation of the algorithm to hide $HI = \{a, b, c\}$.

Transaction number	Items
T1	a, b, c, d
T2	a, b, c
T3	a, c, d

  - Subsets of HI with cardinality $|HI| - 1$: $\{a, b\}$, $\{b, c\}$, $\{a, c\}$.

Data Privacy

Result-driven for association rules mining: heuristic algorithm

- **Example.** Computation of the algorithm to hide $HI = \{a, b, c\}$.

Transaction number	Items
T1	a, b, c, d
T2	a, b, c
T3	a, c, d

  - Subsets of HI with cardinality $|HI| - 1$: $\{a, b\}$, $\{b, c\}$, $\{a, c\}$.
  - $Support(\{a, b\}) = Support(\{b, c\}) = 2$, and $Support(\{a, c\}) = 3$

Data Privacy

Result-driven for association rules mining: heuristic algorithm

- **Example.** Computation of the algorithm to hide $HI = \{a, b, c\}$.

Transaction number	Items
T1	a, b, c, d
T2	a, b, c
T3	a, c, d

  - Subsets of HI with cardinality $|HI| - 1$: $\{a, b\}$, $\{b, c\}$, $\{a, c\}$.
  - $Support(\{a, b\}) = Support(\{b, c\}) = 2$, and $Support(\{a, c\}) = 3$
    $\rightarrow$ We select $HI' = \{a, c\}$.

Data Privacy

Result-driven for association rules mining: heuristic algorithm

- **Example.** Computation of the algorithm to hide $HI = \{a, b, c\}$.

Transaction number	Items
T1	a, b, c, d
T2	a, b, c
T3	a, c, d

  - Subsets of HI with cardinality $|HI| - 1$: $\{a, b\}$, $\{b, c\}$, $\{a, c\}$.
  - $Support(\{a, b\}) = Support(\{b, c\}) = 2$, and $Support(\{a, c\}) = 3$ $\rightarrow$ We select $HI' = \{a, c\}$.
  - Set of transactions in T that support HI (and HI'): $\{T1, T2\}$.

Data Privacy

Result-driven for association rules mining: heuristic algorithm

- **Example.** Computation of the algorithm to hide $HI = \{a, b, c\}$.

Transaction number	Items
T1	a, b, c, d
T2	a, b, c
T3	a, c, d

  - Subsets of HI with cardinality $|HI| - 1$: $\{a, b\}$, $\{b, c\}$, $\{a, c\}$.
  - $Support(\{a, b\}) = Support(\{b, c\}) = 2$, and $Support(\{a, c\}) = 3$ $\rightarrow$ We select $HI' = \{a, c\}$.
  - Set of transactions in T that support HI (and HI'): $\{T1, T2\}$.
  - Ts transaction in $\{T1, T2\}$ that affects the minimum number of itemsets of cardinality 2: $T2$ affects less itemsets than $T1$.

Data Privacy

Result-driven for association rules mining: heuristic algorithm

- **Example.** Computation of the algorithm to hide $HI = \{a, b, c\}$.
- Remove one of the items in $HI' = \{a, c\}$ that are in $T2$:

Data Privacy

Result-driven for association rules mining: heuristic algorithm

- **Example.** Computation of the algorithm to hide $HI = \{a, b, c\}$.
- Remove one of the items in $HI' = \{a, c\}$ that are in $T2$:
  Both have the same support, we select one of them at random.
- Propagate the results forward: recompute supports

Tabular data (Ch. 3.6)

Tabular data

- Aggregates of data with respect to a few variables. 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

Cell (M_2, P_3): number of people with profession P_3 living in municipality M_2.

	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

Cell (M_2, P_3): total salary received by people with profession P_3 living in M_2.

Tabular data

- Aggregates of data do not avoid disclosure
  - **External attack.** Combining the information of the two tables the adversary is able to infer some sensitive information.
    $\Rightarrow (M_2, P_3)$

Tabular data

- Aggregates of data do not avoid disclosure
  - **External attack.** Combining the information of the two tables the adversary is able to infer some sensitive information.
    $\Rightarrow (M_2, P_3)$
  - **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 doctor infers the salary of another doctor.
    $\Rightarrow (M_1, P_1)$

Tabular data

- Aggregates of data do not avoid disclosure
  - **External attack.** Combining the information of the two tables the adversary is able to infer some sensitive information.
    $\Rightarrow (M_2, P_3)$
  - **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 doctor infers the salary of another doctor.
    $\Rightarrow (M_1, P_1)$
  - **Internal attack with dominance.** This is an internal attack where a contribution of one person, say p_0, in a cell is so high that permits p_0 to obtain accurate bounds of the contribution of the others.
    $\Rightarrow (M_3, P_5)$ with 5 people. $salary(p_0) = 350$, then the salary of the other four is at most $363 - 350 = 13$.

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

- **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: data protection mechanism

- Protection of a tabular data
  - Perturbative
    - ⋆ Post-tabular
      - – Rounding
      - – Controlled tabular adjustment (CTA)
    - ⋆ Pre-tabular
  - 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:

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

- Solutions build using optimization

Tabular data: information loss

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