Other Privacy Definitions: l-diversity and t-closeness

Murat Kantarcioglu

Outline

- In this lecture, we will discuss additional privacy definitions that tries to address the limitations of k-anonymity
  – L-diversity
  – T-closeness

L-diversity: Privacy beyond k-anonymity

Following Slides are Based on Machanavajjhala et al., 2006

k-Anonymity

- Each released record should be indistinguishable from at least (k-1) others on its QI attributes
- Alternatively: cardinality of any query result on released data should be at least k
- k-anonymity is (the first) one of many privacy definitions in this line of work
  - l-diversity, t-closeness, m-invariance, delta-presence...

Attacks Against K-Anonymity

- Complementary Release Attack
  - Different releases can be linked together to compromise k-anonymity.
  - Solution:
    - Consider all of the released tables before release the new one, and try to avoid linking.
    - Other data holders may release some data that can be used in this kind of attack. Generally, this kind of attack is hard to be prohibited completely.

Attacks Against K-Anonymity

- **k-Anonymity does not provide privacy if:**
  - **Sensitive values in an equivalence class lack diversity**
  - **The attacker has background knowledge**

Homogeneity Attack

Bob	
Zipcode	***Age***
47678	27

A 3-anonymous patient table

Zipcode	Age	Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
4790*	≥40	Flu
4790*	≥40	Heart Disease
4790*	≥40	Cancer
476**	3*	Heart Disease
476**	3*	Cancer
476**	3*	Cancer

Background Knowledge Attack

Umeko (Japanese)	
Zipcode	***Age***
47673	36

Goals for Privacy-preserving Data Publishing Definitions

- Easy to understand.
- Should prevent background knowledge attacks.
- Should be easily enforceble.

L-diversity principles

- L-diversity principle: A q-block is l-diverse if contains at least l ‘well represented” values for the sensitive attribute S. A table is l-diverse if every q-block is l-diverse

l-Diversity

- Distinct *l*-diversity
  - Each equivalence class has at least ***l*** well-represented sensitive values
  - Limitation:
    - Doesn't prevent the probabilistic inference attacks
    - Ex.

      In one equivalent class, there are ten tuples. In the "Disease" area, one of them is "Cancer", one is "Heart Disease" and the remaining eight are "Flu". This satisfies 3-diversity, but the attacker can still affirm that the target person's disease is "Flu" with the accuracy of 80%.

l-Diversity(Cont'd)

- Entropy *l*-diversity
  - Each equivalence class not only must have enough different sensitive values, but also the different sensitive values must be distributed evenly enough.
  - It means the entropy of the distribution of sensitive values in each equivalence class is at least log(l)
  - Sometimes this maybe too restrictive. When some values are very common, the entropy of the entire table may be very low. This leads to the less conservative notion of l-diversity.

l-Diversity(Cont'd)

- Recursive *(c,l)*-diversity
  - The most frequent value does not appear too frequently
  - $r_1 < c(r_l + r_{l+1} + \ldots + r_m)$

Limitations of *l*-Diversity

l-diversity may be difficult and unnecessary to achieve.

- **A single sensitive attribute**
  - **Two values: HIV positive (1%) and HIV negative (99%)**
  - **Very different degrees of sensitivity**
- **l-diversity is unnecessary to achieve**
  - **2-diversity is unnecessary for an equivalence class that contains only negative records**
- **l-diversity is difficult to achieve**
  - **Suppose there are 10000 records in total**
  - **To have distinct 2-diversity, there can be at most 10000*1%=100 equivalence classes**

Limitations of *l*-Diversity(Cont'd)

l-diversity is insufficient to prevent attribute disclosure.

Similarity Attack

Bob	
Zip	***Age***
47678	27

A 3-diverse patient table

Zipcode	Age	Salary	Disease
476**	2*	20K	Gastric Ulcer
476**	2*	30K	Gastritis
476**	2*	40K	Stomach Cancer
4790*	≥40	50K	Gastritis
4790*	≥40	100K	Flu
4790*	≥40	70K	Bronchitis
476**	3*	60K	Bronchitis
476**	3*	80K	Pneumonia
476**	3*	90K	Stomach Cancer

Conclusion

1. **Bob's salary is in [20k,40k], which is relative low.**
2. **Bob has some stomach-related disease.**

l-diversity does not consider semantic meanings of sensitive values

t-Closeness: Privacy Beyond k-Anonymity and l-Diversity

Based on Li et al., 2007

t-closeness

- k-anonymity prevents identity disclosure but not attribute disclosure
- To solve that problem l-diversity requires that each eq. class has at least l values for each sensitive attribute
- But l-diversity has some limitations
- t-closeness requires that the distribution of a sensitive attribute in any eq. class is close to the distribution of a sensitive attribute in the overall table.

t-closeness: A New Privacy Measure

- Privacy is measured by the information gain of an observer.
- Information Gain = Posterior Belief – Prior Belief
- Q = the distribution of the sensitive attribute in the whole table
- P = the distribution of the sensitive attribute in eq. class

t-closeness Principle

- An equivalence class is said to have t-closeness
  - if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t.
- A table is said to have t-closeness
  - if all equivalence classes have t-closeness.

Measuring the distance between two probabilistic distributions

- Given two distributions

  $P = (p_1, p_2, \ldots, p_m)$, $Q = (q_1, q_2, \ldots, q_m)$,

  two well-known distance measures are as follows.

  The variational distance is defined as:

$$\mathrm{D}[\mathbf{P}, \mathbf{Q}] = \sum_{i=1}^{m} \frac{1}{2} |p_i - q_i|.$$

Earth Mover's Distance

$$WORK(\mathbf{P}, \mathbf{Q}, F) = \sum_{i=1}^{m} \sum_{j=1}^{m} d_{ij} f_{ij}$$

subject to the following constraints:

$$f_{ij} \geq 0 \qquad 1 \leq i \leq m, 1 \leq j \leq m \qquad (c1)$$

$$p_i - \sum_{j=1}^{m} f_{ij} + \sum_{j=1}^{m} f_{ji} = q_i \qquad 1 \leq i \leq m \qquad (c2)$$

$$\sum_{i=1}^{m} \sum_{j=1}^{m} f_{ij} = \sum_{i=1}^{m} p_i = \sum_{i=1}^{m} q_i = 1 \qquad (c3)$$

Earth Mover's Distance

These three constraints guarantee that $\mathbf{P}$ is transformed to $\mathbf{Q}$ by the mass flow F. Once the transportation problem is solved, the EMD is defined to be the total work,[3] i.e.,

$$\mathrm{D}[\mathbf{P}, \mathbf{Q}] = WORK(\mathbf{P}, \mathbf{Q}, F) = \sum_{i=1}^{m} \sum_{j=1}^{m} d_{ij} f_{ij}$$

Similarity Attack Example

	ZIP Code	Age	Salary	Disease
1	4767*	≤ 40	3K	gastric ulcer
3	4767*	≤ 40	5K	stomach cancer
8	4767*	≤ 40	9K	pneumonia
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
2	4760*	≤ 40	4K	gastritis
7	4760*	≤ 40	7K	bronchitis
9	4760*	≤ 40	10K	stomach cancer

Table 5. Table that has 0.167-closeness w.r.t. Salary and 0.278-closeness w.r.t. Disease

Conclusion

- t-closeness protects against attribute disclosure but not identity disclosure
- t-closeness requires that the distribution of a sensitive attribute in any eq. class is close to the distribution of a sensitive attribute in the overall table.