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**

<table>
<thead>
<tr>
<th>Bob</th>
<th>Zipcode</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>47678</td>
<td>27</td>
</tr>
</tbody>
</table>

**Background Knowledge Attack**

<table>
<thead>
<tr>
<th>Umeko (Japanese)</th>
<th>Zipcode</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>47673</td>
<td>36</td>
</tr>
</tbody>
</table>

A 3-anonymous patient table:

<table>
<thead>
<tr>
<th>Zipcode</th>
<th>Age</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>476**</td>
<td>2*</td>
<td>Heart Disease</td>
</tr>
<tr>
<td>476**</td>
<td>2*</td>
<td>Heart Disease</td>
</tr>
<tr>
<td>476**</td>
<td>2*</td>
<td>Heart Disease</td>
</tr>
<tr>
<td>4790*</td>
<td>≥40</td>
<td>Flu</td>
</tr>
<tr>
<td>4790*</td>
<td>≥40</td>
<td>Heart Disease</td>
</tr>
<tr>
<td>4790*</td>
<td>≥40</td>
<td>Cancer</td>
</tr>
<tr>
<td>476**</td>
<td>3*</td>
<td>Heart Disease</td>
</tr>
<tr>
<td>476**</td>
<td>3*</td>
<td>Cancer</td>
</tr>
<tr>
<td>476**</td>
<td>3*</td>
<td>Cancer</td>
</tr>
</tbody>
</table>
Goals for Privacy-preserving Data Publishing Definitions

- Easy to understand.
- Should prevent background knowledge attacks.
- Should be easily enforceable.
L-diversity principles

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

- Distinct /-diversity
  - Each equivalence class has at least / 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%.
- 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.
• 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**

<table>
<thead>
<tr>
<th>Zip</th>
<th>Age</th>
<th>Salary</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>476**</td>
<td>2*</td>
<td>20K</td>
<td>Gastric Ulcer</td>
</tr>
<tr>
<td>476**</td>
<td>2*</td>
<td>30K</td>
<td>Gastritis</td>
</tr>
<tr>
<td>476**</td>
<td>2*</td>
<td>40K</td>
<td>Stomach Cancer</td>
</tr>
<tr>
<td>4790*</td>
<td>≥40</td>
<td>50K</td>
<td>Gastritis</td>
</tr>
<tr>
<td>4790*</td>
<td>≥40</td>
<td>100K</td>
<td>Flu</td>
</tr>
<tr>
<td>4790*</td>
<td>≥40</td>
<td>70K</td>
<td>Bronchitis</td>
</tr>
<tr>
<td>476**</td>
<td>3*</td>
<td>60K</td>
<td>Bronchitis</td>
</tr>
<tr>
<td>476**</td>
<td>3*</td>
<td>80K</td>
<td>Pneumonia</td>
</tr>
<tr>
<td>476**</td>
<td>3*</td>
<td>90K</td>
<td>Stomach Cancer</td>
</tr>
</tbody>
</table>

**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.
Given two distributions

\[ P = (p_1, p_2, ..., p_m), \quad Q = (q_1, q_2, ..., q_m), \]

two well-known distance measures are as follows. The variational distance is defined as:

\[
D[P, Q] = \sum_{i=1}^{m} \frac{1}{2} |p_i - q_i|.
\]
Earth Mover’s Distance

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

subject to the following constraints:

\[ f_{ij} \geq 0 \quad 1 \leq i \leq m, 1 \leq j \leq m \quad (c1) \]

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

\[ \sum_{i=1}^{m} \sum_{j=1}^{m} f_{ij} = \sum_{i=1}^{m} p_i = \sum_{i=1}^{m} q_i = 1 \quad (c3) \]
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, i.e.,

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

<table>
<thead>
<tr>
<th></th>
<th>ZIP Code</th>
<th>Age</th>
<th>Salary</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4767*</td>
<td>≤ 40</td>
<td>3K</td>
<td>gastric ulcer</td>
</tr>
<tr>
<td>3</td>
<td>4767*</td>
<td>≤ 40</td>
<td>5K</td>
<td>stomach cancer</td>
</tr>
<tr>
<td>8</td>
<td>4767*</td>
<td>≤ 40</td>
<td>9K</td>
<td>pneumonia</td>
</tr>
<tr>
<td>4</td>
<td>4790*</td>
<td>≥ 40</td>
<td>6K</td>
<td>gastritis</td>
</tr>
<tr>
<td>5</td>
<td>4790*</td>
<td>≥ 40</td>
<td>11K</td>
<td>flu</td>
</tr>
<tr>
<td>6</td>
<td>4790*</td>
<td>≥ 40</td>
<td>8K</td>
<td>bronchitis</td>
</tr>
<tr>
<td>2</td>
<td>4760*</td>
<td>≤ 40</td>
<td>4K</td>
<td>gastritis</td>
</tr>
<tr>
<td>7</td>
<td>4760*</td>
<td>≤ 40</td>
<td>7K</td>
<td>bronchitis</td>
</tr>
<tr>
<td>9</td>
<td>4760*</td>
<td>≤ 40</td>
<td>10K</td>
<td>stomach cancer</td>
</tr>
</tbody>
</table>

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.