CS573 Data Privacy and Security

Anonymization methods

Li Xiong
Today

• Clustering based anonymization (cont)
• Permutation based anonymization
• Other privacy principles
Microaggregation/Clustering

• Two steps:
  – Partition original dataset into clusters of similar records containing at least $k$ records
  – For each cluster, compute an aggregation operation and use it to replace the original records
    • e.g., mean for continuous data, median for categorical data
What is Clustering?

- Finding groups of objects (clusters)
  - Objects similar to one another in the same group
  - Objects different from the objects in other groups
- Unsupervised learning

Intra-cluster distances are minimized

Inter-cluster distances are maximized
Clustering Approaches

• **Partitioning approach:**
  – Construct various partitions and then evaluate them by some criterion, e.g.,
    minimizing the sum of square errors
  – Typical methods: k-means, k-medoids, CLARANS

• **Hierarchical approach:**
  – Create a hierarchical decomposition of the set of data (or objects) using some criterion
  – Typical methods: Diana, Agnes, BIRCH, ROCK, CAMELEON

• **Density-based approach:**
  – Based on connectivity and density functions
  – Typical methods: DBSACN, OPTICS, DenClue

• **Others**
K-Means Clustering: Lloyd Algorithm

• Given $k$, and randomly choose $k$ initial cluster centers
• Partition objects into $k$ nonempty subsets by assigning each object to the cluster with the nearest centroid
• Update centroid, i.e. mean point of the cluster
• Go back to Step 2, stop when no more new assignment
The *K-Means* Clustering Method

- **Example**

  K=2

  Arbitrarily choose K object as initial cluster center

  Assign each objects to most similar center

  Update the cluster means

  reassign

  Update the cluster means

  reassign
Hierarchical Clustering

- Produces a set of nested clusters organized as a hierarchical tree
- Can be visualized as a dendrogram
  - A tree like diagram representing a hierarchy of nested clusters
  - Clustering obtained by cutting at desired level
Hierarchical Clustering

• Two main types of hierarchical clustering
  – Agglomerative:
    • Start with the points as individual clusters
    • At each step, merge the closest pair of clusters until only one cluster (or k clusters) left
  – Divisive:
    • Start with one, all-inclusive cluster
    • At each step, split a cluster until each cluster contains a point (or there are k clusters)
Agglomerative Clustering Algorithm

1. Compute the proximity matrix
2. Let each data point be a cluster
3. Repeat
4. Merge the two closest clusters
5. Update the proximity matrix
6. Until only a single cluster remains
Starting Situation

- Start with clusters of individual points and a proximity matrix
Intermediate Situation

Proximity Matrix
How to Define Inter-Cluster Similarity

Similarity?

Proximity Matrix

<table>
<thead>
<tr>
<th></th>
<th>p1</th>
<th>p2</th>
<th>p3</th>
<th>p4</th>
<th>p5</th>
<th>. . .</th>
</tr>
</thead>
<tbody>
<tr>
<td>p1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Distance Between Clusters

- **Single Link**: smallest distance between points
- **Complete Link**: largest distance between points
- **Average Link**: average distance between points
- **Centroid**: distance between centroids
Clustering for Anonymization

• Are they directly applicable?
• Which algorithms are directly applicable?
  – K-means; hierarchical
Anonymization And Clustering

• *k*-Member Clustering Problem
  – From a given set of *n* records, find a set of clusters such that
    • Each cluster contains at least *k* records, and
    • The total intra-cluster distance is minimized.
  – The problem is NP-complete
Anonymization using Microaggregation or Clustering

- Practical Data-Oriented Microaggregation for Statistical Disclosure Control, Domingo-Ferrer, TKDE 2002
- Ordinal, Continuous and Heterogeneous k-anonymity through microaggregation, Domingo-Ferrer, DMKD 2005
- Achieving anonymity via clustering, Aggarwal, PODS 2006
- Efficient k-anonymization using clustering techniques, Byun, DASFAA 2007
Multivariate microaggregation algorithm

Basic idea:

- Form two k-member clusters at each step
- Form one k-member cluster for remaining records, if available
- Form one cluster for remaining records
Multivariate microaggregation algorithm  
(Maximum Distance to Average Vector)  
MDAV-generic(R: dataset, k: integer)  
while |R| \geq 3k  
1. compute average record \sim{x} of all records in R  
2. find most distant record \(x_r\) from \sim{x}  
3. find most distant record \(x_s\) from \(x_r\)  
4. form two clusters from k-1 records closest to \(x_r\) and k-1 closest to \(x_s\)  
5. Remove the clusters from R and run MDAV-generic on the remaining dataset  
end while  
if 3k-1 \leq |R| \leq 2k  
1. compute average record \sim{x} of remaining records in R  
2. find the most distant record \(x_r\) from \sim{x}  
3. form a cluster from k-1 records closest to \sim{x}  
4. form another cluster containing the remaining records  
else (fewer than 2k records in R) form a new cluster from the remaining records
MDAV-generic for continuous attributes

- use arithmetic mean and Euclidean distance
- standardize attributes (subtract mean and divide by standard deviation) to give them equal weight for computing distances
- After MDAV-generic, destandardize attributes
MDAV-generic for categorical attributes

- The distance between two ordinal attributes $a$ and $b$ in an attribute $V_i$:

$$d_{ord}(a,b) = \frac{|\{i| \leq i < b\}|}{|D(V_i)|}$$

- i.e., the number of categories separating $a$ and $b$ divided by the number of categories in the attribute

- The distance between two nominal attributes is defined according to equality: 0 if they're equal, else 1
Empirical Results

• Continuous attributes
    • 1080 records described by 13 continuous attributes
    • Computed $k$-anonymity for $k = 3, \ldots, 9$ and quasi-identifiers with 6 and 13 attributes

• Categorical attributes
  – From the U.S. Housing Survey (1993)
    • Three ordinal and eight nominal attributes
    • Computed $k$-anonymity for $k = 2, \ldots, 9$ and quasi-identifiers with 3, 4, 8 and 11 attributes
IL measures for continuous attributes

- IL1 = mean variation of individual attributes in original and k-anonymous datasets
- IL2 = mean variation of attribute means in both datasets
- IL3 = mean variation of attribute variances
- IL4 = mean variation of attribute covariances
- IL5 = mean variation of attribute Pearson's correlations
- IL6 = 100 times the average of IL1-6
MDAV-generic preserves means and variances (IL2 and IL3)

The impact on the non-preserved statistics grows with the quasi-identifier length, as one would expect.

For a fixed-quasi-identifier length, the impact on the non-preserved statistics grows with $k$.
Anonymization using Microaggregation or Clustering

- Practical Data-Oriented Microaggregation for Statistical Disclosure Control, Domingo-Ferrer, TKDE 2002
- Ordinal, Continuous and Heterogeneous k-anonymity through microaggregation, Domingo-Ferrer, DMKD 2005
- Achieving anonymity via clustering, Aggarwal, PODS 2006
- Efficient k-anonymization using clustering techniques, Byun, DASFAAA 2007
Greedy Algorithm

• Basic idea:
  – Find k-member clusters, one cluster at a time
  – Assign remaining $<k$ points to the previous clusters

• Some details
  – How to compute distances between records
  – How to find centroid?
  – How to find the best point to join current cluster?
Distance between two categorical values

- Equally different to each other.
  - 0 if they are the same
  - 1 if they are different

- Relationships can be easily captured in a taxonomy tree.
Distance between two categorical values

• **Definition**

Let $D$ be a categorical domain and $TD$ be a taxonomy tree defined for $D$. The normalized distance between two values $v_i, v_j \in D$ is defined as:

$$\delta_C(v_1, v_2) = \frac{H(\wedge(v_i, v_j))}{H(T_D)}$$

where $\wedge(x, y)$ is the subtree rooted at the lowest common ancestor of $x$ and $y$, and $H(T)$ represents the height of tree $T$.

Example:

The distance between India and USA is $3/3 = 1$.

The distance between India and Iran is $2/3 = 0.66$. 

![Taxonomy tree of Country](image-url)
Cost Function - Information loss (IL)

- The amount of distortion (i.e., information loss) caused by the generalization process.
  
  Note: Records in each cluster are generalized to share the same quasi-identifier value that represents every original quasi-identifier value in the cluster.

- Definition: Let \( e = \{r_1, \ldots, r_k\} \) be a cluster (i.e., equivalence class). Then the amount of information loss in \( e \), denoted by \( IL(e) \), is defined as:

\[
IL(e) = |e| \cdot D(e)
\]

\[
D(e) = \sum_{i=1,\ldots,m} \frac{(MAX_{N_i} - MIN_{N_i})}{|N_i|} + \sum_{j=1,\ldots,n} \frac{H(\Lambda(\cup C_j))}{H(T_{C_j})}
\]

where \( |e| \) is the number of records in \( e \), \(|N|\) represents the size of numeric domain \( N \), \( \Lambda(\cup C_i) \) is the subtree rooted at the lowest common ancestor of every value in \( \cup C_i \), and \( H(T) \) is the height of tree \( T \).
Cost Function - Information loss (IL)

Example

<table>
<thead>
<tr>
<th></th>
<th>Age</th>
<th>Country</th>
<th>Occupation</th>
<th>Salary</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>r1</td>
<td>41</td>
<td>USA</td>
<td>Armed-Forces</td>
<td>≥50K</td>
<td>Cancer</td>
</tr>
<tr>
<td>r2</td>
<td>57</td>
<td>India</td>
<td>Tech-support</td>
<td>&lt;50K</td>
<td>Flu</td>
</tr>
<tr>
<td>r3</td>
<td>40</td>
<td>Canada</td>
<td>Teacher</td>
<td>&lt;50K</td>
<td>Obesity</td>
</tr>
<tr>
<td>r4</td>
<td>38</td>
<td>Iran</td>
<td>Tech-support</td>
<td>≥50K</td>
<td>Flu</td>
</tr>
<tr>
<td>r5</td>
<td>24</td>
<td>Brazil</td>
<td>Doctor</td>
<td>≥50K</td>
<td>Cancer</td>
</tr>
<tr>
<td>r6</td>
<td>45</td>
<td>Greece</td>
<td>Salesman</td>
<td>&lt;50K</td>
<td>Fever</td>
</tr>
</tbody>
</table>

Cluster e₁

<table>
<thead>
<tr>
<th>Age</th>
<th>Country</th>
<th>Occupation</th>
<th>Salary</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>41</td>
<td>USA</td>
<td>Armed-Forces</td>
<td>≥50K</td>
<td>Cancer</td>
</tr>
<tr>
<td>40</td>
<td>Canada</td>
<td>Teacher</td>
<td>&lt;50K</td>
<td>Obesity</td>
</tr>
<tr>
<td>24</td>
<td>Brazil</td>
<td>Doctor</td>
<td>≥50K</td>
<td>Cancer</td>
</tr>
</tbody>
</table>

Cluster e₂

<table>
<thead>
<tr>
<th>Age</th>
<th>Country</th>
<th>Occupation</th>
<th>Salary</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>41</td>
<td>USA</td>
<td>Armed-Forces</td>
<td>≥50K</td>
<td>Cancer</td>
</tr>
<tr>
<td>57</td>
<td>India</td>
<td>Tech-support</td>
<td>&lt;50K</td>
<td>Flu</td>
</tr>
<tr>
<td>24</td>
<td>Brazil</td>
<td>Doctor</td>
<td>≥50K</td>
<td>Cancer</td>
</tr>
</tbody>
</table>

IL(e₁) = 3 • D(e₁)  
D(e₁) = (41-24)/33 + (2/3) + 1 = 2.1818...  
IL(e₁) = 3 • 2.1818... = 6.5454...

IL(e₂) = 3 • D(e₂)  
D(e₂) = (57-24)/33 + (3/3) + 1 = 3  
IL(e₂) = 3 • 3 = 9
Greedy k-member clustering algorithm

Function \texttt{greedy\_k\_member\_clustering} (S, k)
Input: a set of records S and a threshold value k.
Output: a set of clusters each of which contains at least k records.
1. if (|S| \leq k)
2. \quad \text{return } S;
3. \quad \text{end if;}
4. \text{result } = \emptyset;
5. \quad r = \text{a randomly picked record from } S;
6. \quad \text{while (|S| \geq k)}
7. \quad \quad r = \text{the furthest record from } r;
8. \quad \quad S = S \setminus \{r\};
9. \quad \quad c = \{r\};
10. \quad \quad \text{while (|c| < k)}
11. \quad \quad \quad r = \text{find\_best\_record}(S, c);
12. \quad \quad \quad S = S \setminus \{r\};
13. \quad \quad \quad c = c \cup \{r\};
14. \quad \quad \text{end while;}
15. \quad \text{result } = \text{result } \cup \{c\};
16. \text{end while;}
17. \text{while (|S| \neq 0)}
18. \quad r = \text{a randomly picked record from } S;
19. \quad S = S \setminus \{r\};
20. \quad c = \text{find\_best\_cluster}(\text{result }, r);
21. \quad c = c \cup \{r\};
22. \text{end while;}
23. \text{return result;}

Function \texttt{find\_best\_record} (S, c)
Input: a set of records S and a cluster c.
Output: a record r \in S such that IL(c \cup \{r\}) is minimal.
1. \quad n = |S|;
2. \quad \min = \infty;
3. \quad \best = \text{null};
4. \quad \text{for}(i = 1, \ldots n)
5. \quad \quad r = \text{-i-th record in } S;
6. \quad \quad \text{diff } = \text{IL}(c \cup \{r\}) - \text{IL}(c);
7. \quad \quad \text{if (diff < min)}
8. \quad \quad \quad \min = \text{diff};
9. \quad \quad \quad \best = r;
10. \quad \text{end if;}
11. \text{end for;}
12. \text{return best;}

Function \texttt{find\_best\_cluster} (C, r)
Input: a set of clusters C and a record r.
Output: a cluster c \in C such that IL(c \cup \{r\}) is minimal.
1. \quad n = |C|;
2. \quad \min = \infty;
3. \quad \best = \text{null};
4. \quad \text{for}(i = 1, \ldots n)
5. \quad \quad c = \text{i-th cluster in } C;
6. \quad \quad \text{diff } = \text{IL}(c \cup \{r\}) - \text{IL}(c);
7. \quad \quad \text{if (diff < min)}
8. \quad \quad \quad \min = \text{diff};
9. \quad \quad \quad \best = c;
10. \quad \text{end if;}
11. \text{end for;}
12. \text{return best;}

End;
classification metric (CM)

– preserve the correlation between quasi-identifier and class labels (non-sensitive values)

\[ CM = \sum_{all \ rows} Penalty(\ row \ r) / N \]

Where \( N \) is the total number of records, and \( Penalty(\ row \ r) = 1 \) if \( r \) is suppressed or the class label of \( r \) is different from the class label of the majority in the equivalence group.
Experimentl Results

• Experimental Setup
  – Data: Adult dataset from the UC Irvine Machine Learning Repository
    • 10 attributes (2 numeric, 7 categorical, 1 class)
  – Compare with 2 other algorithms
    • Median partitioning (Mondrian algorithm)
    • $k$-Nearest neighbor
Experimentl Results
Conclusion

• Transforming the k-anonymity problem to the k-member clustering problem
• Overall the Greedy Algorithm produced better results compared to other algorithms at the cost of efficiency
Today

• Clustering based anonymization (cont)
• Permutation based anonymization
• Other privacy principles
Anonymization methods

• Non-perturbative: don't distort the data
  – Generalization
  – Suppression

• Perturbative: distort the data
  – Microaggregation/clustering
  – Additive noise

• Anatomization and permutation
  – De-associate relationship between QID and sensitive attribute
Problems with Generalization and Clustering

<table>
<thead>
<tr>
<th>tuple ID</th>
<th>Age</th>
<th>Sex</th>
<th>Zipcode</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Bob)</td>
<td>23</td>
<td>M</td>
<td>11000</td>
<td>pneumonia</td>
</tr>
<tr>
<td>2</td>
<td>27</td>
<td>M</td>
<td>13000</td>
<td>Dyspepsia</td>
</tr>
<tr>
<td>3</td>
<td>35</td>
<td>M</td>
<td>59000</td>
<td>Dyspepsia</td>
</tr>
<tr>
<td>4</td>
<td>59</td>
<td>M</td>
<td>12000</td>
<td>pneumonia</td>
</tr>
<tr>
<td>5</td>
<td>61</td>
<td>F</td>
<td>54000</td>
<td>flu</td>
</tr>
<tr>
<td>6</td>
<td>65</td>
<td>F</td>
<td>25000</td>
<td>stomach pain</td>
</tr>
<tr>
<td>7 (Alice)</td>
<td>65</td>
<td>F</td>
<td>25000</td>
<td>flu</td>
</tr>
<tr>
<td>8</td>
<td>70</td>
<td>F</td>
<td>30000</td>
<td>bronchitis</td>
</tr>
</tbody>
</table>

Query A:

```
SELECT COUNT(*)
FROM Microdata
WHERE Disease = 'pneumonia'
AND Age <= 30
AND Zipcode IN [10001, 20000]
```
Querying generalized table

- $R_1$ and $R_2$ are the anonymized QID groups
- $Q$ is the query range
- **Uniform assumption**
  \[
  p = \frac{\text{Area}(R_1 \cap R_Q)}{\text{Area}(R_1)} = \frac{10 \times 10}{50 \times 40} = 0.05
  \]
- Estimated Answer for A: $4(0.05) = 0.2$
Concept of the Anatomy Algorithm

• Release 2 tables, **quasi-identifier table (QIT)** and **sensitive table (ST)**

• Use the same QI groups (satisfy 1-diversity), replace the sensitive attribute values with a Group-ID column

• Then produce a sensitive table with *Disease* statistics

<table>
<thead>
<tr>
<th>tuple ID</th>
<th>Age</th>
<th>Sex</th>
<th>Zipcode</th>
<th>Group-ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>23</td>
<td>M</td>
<td>11000</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>27</td>
<td>M</td>
<td>13000</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>35</td>
<td>M</td>
<td>59000</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>59</td>
<td>M</td>
<td>12000</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>61</td>
<td>F</td>
<td>54000</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>65</td>
<td>F</td>
<td>25000</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>65</td>
<td>F</td>
<td>25000</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>70</td>
<td>F</td>
<td>30000</td>
<td>2</td>
</tr>
</tbody>
</table>

**QIT**

<table>
<thead>
<tr>
<th>Group-ID</th>
<th>Disease</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>headache</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>pneumonia</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>bronchitis</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>flu</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>stomach ache</td>
<td>1</td>
</tr>
</tbody>
</table>

**ST**
Concept of the Anatomy Algorithm

<table>
<thead>
<tr>
<th>Tuple ID</th>
<th>Age</th>
<th>Sex</th>
<th>Zipcode</th>
<th>Group-ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>23</td>
<td>M</td>
<td>11000</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>27</td>
<td>M</td>
<td>13000</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>35</td>
<td>M</td>
<td>59000</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>59</td>
<td>M</td>
<td>12000</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>61</td>
<td>F</td>
<td>54000</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>65</td>
<td>F</td>
<td>25000</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>65</td>
<td>F</td>
<td>25000</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>70</td>
<td>F</td>
<td>30000</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Group-ID</th>
<th>Disease</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>headache</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>pneumonia</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>bronchitis</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>flu</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>stomach ache</td>
<td>1</td>
</tr>
</tbody>
</table>

QIT

• Does it satisfy k-anonymity? l-diversity?

• Query results?

```sql
SELECT COUNT(*)
FROM Microdata
WHERE Disease = 'pneumonia' AND Age <= 30 AND Zipcode IN [10001,20000]
```
DEFINITION 3. (Anatomy)
With a given $l$-diverse partition anatomy will create QIT and ST tables
QIT will be constructed as the following:
$$(A^{q_1}, A^{q_2}, ..., A^{q_d}, \text{Group-ID})$$

ST will be constructed as the following:
$$(\text{Group-ID}, A^s, \text{Count})$$
Privacy properties

**THEOREM 1.** Given a pair of QIT and ST inference of the sensitive value of any individual is at most $1/l$
Comparison with generalization

- Compare with generalization on two assumptions:
  A1: the adversary has the QI-values of the target individual
  A2: the adversary also knows that the individual is definitely in the microdata

If A1 and A2 are true, anatomy is as good as generalization $\frac{1}{l}$ holds true
If A1 is true and A2 is false, generalization is stronger
If A1 and A2 are false, generalization is still stronger
Preserving Data Correlation

• Examine the correlation between Age and Disease in T using probability density function pdf

• Example: $t_1$

<table>
<thead>
<tr>
<th>tuple ID</th>
<th>Age</th>
<th>Sex</th>
<th>Zipcode</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Bob)</td>
<td>23</td>
<td>M</td>
<td>11000</td>
<td>pneumonia</td>
</tr>
<tr>
<td>2</td>
<td>27</td>
<td>M</td>
<td>13000</td>
<td>Dyspepsia</td>
</tr>
<tr>
<td>3</td>
<td>35</td>
<td>M</td>
<td>59000</td>
<td>Dyspepsia</td>
</tr>
<tr>
<td>4</td>
<td>59</td>
<td>M</td>
<td>12000</td>
<td>pneumonia</td>
</tr>
<tr>
<td>5</td>
<td>61</td>
<td>F</td>
<td>54000</td>
<td>flu</td>
</tr>
<tr>
<td>6</td>
<td>65</td>
<td>F</td>
<td>25000</td>
<td>stomach pain</td>
</tr>
<tr>
<td>7 (Alice)</td>
<td>65</td>
<td>F</td>
<td>25000</td>
<td>flu</td>
</tr>
<tr>
<td>8</td>
<td>70</td>
<td>F</td>
<td>30000</td>
<td>bronchitis</td>
</tr>
</tbody>
</table>

table 1

\[ G_{t_1}(x) = \begin{cases} 1 & \text{if } x = (t_1[A], t_1[D]) \\ 0 & \text{otherwise} \end{cases} \]
Preserving Data Correlation \textit{cont.}

- To re-construct an approximate pdf of $t_1$ from the generalization table:

<table>
<thead>
<tr>
<th>tuple ID</th>
<th>Age</th>
<th>Sex</th>
<th>Zipcode</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>[21,60]</td>
<td>M</td>
<td>[10001, 60000]</td>
<td>pneumonia</td>
</tr>
<tr>
<td>2</td>
<td>[21,60]</td>
<td>M</td>
<td>[10001, 60000]</td>
<td>Dyspepsia</td>
</tr>
<tr>
<td>3</td>
<td>[21,60]</td>
<td>M</td>
<td>[10001, 60000]</td>
<td>Dyspepsia</td>
</tr>
<tr>
<td>4</td>
<td>[21,60]</td>
<td>M</td>
<td>[10001, 60000]</td>
<td>pneumonia</td>
</tr>
<tr>
<td>5</td>
<td>[61,70]</td>
<td>F</td>
<td>[10001, 60000]</td>
<td>flu</td>
</tr>
<tr>
<td>6</td>
<td>[61,70]</td>
<td>F</td>
<td>[10001, 60000]</td>
<td>stomach pain</td>
</tr>
<tr>
<td>7</td>
<td>[61,70]</td>
<td>F</td>
<td>[10001, 60000]</td>
<td>flu</td>
</tr>
<tr>
<td>8</td>
<td>[61,70]</td>
<td>F</td>
<td>[10001, 60000]</td>
<td>bronchitis</td>
</tr>
</tbody>
</table>

(b) Approximated from generalization

$$\tilde{g}_{t_1}^{gen}(x) = \begin{cases} 
1/40 & \text{if } x[A] \in [21, 60] \text{ and } x[D] = \text{pneumonia} \\
0 & \text{otherwise}
\end{cases}$$
Preserving Data Correlation cont.

- To re-construct an approximate pdf of $t_1$ from the QIT and ST tables:

<table>
<thead>
<tr>
<th>tuple ID</th>
<th>Age</th>
<th>Sex</th>
<th>Zipcode</th>
<th>Group-ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>23</td>
<td>M</td>
<td>11000</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>27</td>
<td>M</td>
<td>13000</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>35</td>
<td>M</td>
<td>59000</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>59</td>
<td>M</td>
<td>12000</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>61</td>
<td>F</td>
<td>54000</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>65</td>
<td>F</td>
<td>25000</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>65</td>
<td>F</td>
<td>25000</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>70</td>
<td>F</td>
<td>30000</td>
<td>2</td>
</tr>
</tbody>
</table>

QIT

<table>
<thead>
<tr>
<th>Group-ID</th>
<th>Disease</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>headache</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>pneumonia</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>bronchitis</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>flu</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>stomach ache</td>
<td>1</td>
</tr>
</tbody>
</table>

ST

$$\tilde{\mathcal{G}}_{t_1}^\text{ana}(x) = \begin{cases} 
1/2 & \text{if } x = (23, \text{pneumonia}) \text{ or } x = (23, \text{dyspepsia}) \\
0 & \text{otherwise}
\end{cases}$$
Preserving Data Correlation \textit{cont.}

- To figure out a more rigorous comparison, calculate the \textit{“L}_2\textit{ distance”} with the following equation:

\[
\sum_{x \in D S_{A,D}} \left( \tilde{g}_{t_1}(x) - g_{t_1}(x) \right)^2
\]

The distance for anatomy is 0.5 while the distance for generalization is 22.5
Preserving Data Correlation cont.

Idea: Measure the error for each pdf by using the following formula:

$$Err_t = \int_{x \in \mathcal{D}} \left( \tilde{G}_t(x) - G_t(x) \right)^2 \, dx$$

Objective: for all tuples $t$ in $T$ and obtain a minimal reconstruction error ($\forall t \in T$)

$$RCE = \sum_{\forall t \in T} Err_t$$

Algorithm: Nearly-Optimal Anatomizing Algorithm
Experiments

- dataset CENSUS that contained the personal information of 500k American adults containing 9 discrete attributes
- Created two sets of microdata tables
  Set 1: 5 tables denoted as OCC-3, ..., OCC-7 so that OCC-$d$ ($3 \leq d \leq 7$) uses the first $d$ as QI-attributes and Occupation as the sensitive attribute $A^s$
  Set 2: 5 tables denoted as SAL-3, ..., SAL-7 so that SAL-$d$ ($3 \leq d \leq 7$) uses the first $d$ as QI-attributes and Salary-class as the sensitive attribute $A^s$
Experiments cont.

Figure 4: Query accuracy vs. the number $d$ of QI-attributes
Today

• Clustering based anonymization (cont)
• Permutation based anonymization
• Other privacy principles