# Data Anonymization Techniques

Anastasia Tugaenko, Viktoria Gingina,
Kira Matveeva, Mikhail Chupilko, Sari Haj Hussein

The Summer School in Software Engineering and Verification (SSSEV 2011)

2011-07-26

- Organizations typically need to publish microdata e.g., census data or election data
- Data publishing gives useful information to researchers and analyzers
- At the same time, it extends a privacy risk to individuals whose data is being published
- We need strong privacy notions that enable us to confine the disclosure risk while simultaneously maximize the benefits

## Skewness Attack

- When the overall distribution is skewed, satisfying l-diversity does not prevent attribute disclosure

- Original data $\rightsquigarrow$ one sensitive attribute $\rightsquigarrow$ test result for a virus (positive or negative)
- Population of 10 000 records $\rightsquigarrow$ 99% negative, 1% positive
- In one EC $\rightsquigarrow$ positive records = negative records $\rightsquigarrow$ 2-diversity
- Privacy risk $\rightsquigarrow$ anyone in this EC has 50% possibility of being positive

## Skewness Attack

- When the overall distribution is skewed, satisfying l-diversity does not prevent attribute disclosure

- Original data $\rightsquigarrow$ one sensitive attribute $\rightsquigarrow$ test result for a virus (positive or negative)
- Population of 10 000 records $\rightsquigarrow$ 99% negative, 1% positive
- In one EC $\rightsquigarrow$ positive records = negative records $\rightsquigarrow$ 2-diversity
- Privacy risk $\rightsquigarrow$ anyone in this EC has 50% possibility of being positive

## Skewness Attack

- When the overall distribution is skewed, satisfying l-diversity does not prevent attribute disclosure

- Original data $\rightsquigarrow$ one sensitive attribute $\rightsquigarrow$ test result for a virus (positive or negative)
- Population of 10 000 records $\rightsquigarrow$ 99% negative, 1% positive
- In one EC $\rightsquigarrow$ positive records $=$ negative records $\rightsquigarrow$ 2-diversity
- Privacy risk $\rightsquigarrow$ anyone in this EC has 50% possibility of being positive

## Similarity Attack

- When the sensitive attribute values in an EC are distinct but semantically similar, an adversary can learn important information

| | ZIP Code | Age | Salary | Disease |
|---|---|---|---|---|
| 1 | 47677 | 29 | 3K | gastric ulcer |
| 2 | 47602 | 22 | 4K | gastritis |
| 3 | 47678 | 27 | 5K | stomach cancer |
| 4 | 47905 | 43 | 6K | gastritis |
| 5 | 47909 | 52 | 11K | flu |
| 6 | 47906 | 47 | 8K | bronchitis |
| 7 | 47605 | 30 | 7K | bronchitis |
| 8 | 47673 | 36 | 9K | pneumonia |
| 9 | 47607 | 32 | 10K | stomach cancer |

**Table 3. Original Salary/Disease Table**

| | ZIP Code | Age | Salary | Disease |
|---|---|---|---|---|
| 1 | 476** | 2* | 3K | gastric ulcer |
| 2 | 476** | 2* | 4K | gastritis |
| 3 | 476** | 2* | 5K | stomach cancer |
| 4 | 4790* | $\geq 40$ | 6K | gastritis |
| 5 | 4790* | $\geq 40$ | 11K | flu |
| 6 | 4790* | $\geq 40$ | 8K | bronchitis |
| 7 | 476** | 3* | 7K | bronchitis |
| 8 | 476** | 3* | 9K | pneumonia |
| 9 | 476** | 3* | 10K | stomach cancer |

**Table 4. A 3-diverse version of Table 3**

## The t-closeness Privacy Measure

- Privacy is measured by the information gain of an observer
- Before seeing the released table, the observer has a prior belief about the sensitive attribute value
- After seeing the released table, the observer has a posterior belief about the sensitive attribute value
- Information gain = posterior belief - prior belief
- Prior belief is influenced by $Q$ = the distribution of the sensitive attribute value in the whole table
- Posterior belief is influenced by $P$ = the distribution of the sensitive attribute value in the EC
- Information gain = $D[P, Q]$

## The t-closeness Principle

- An EC has t-closeness if $D[P, Q] \leq t$
- A table has t-closeness if all ECs have t-closeness
- $\downarrow D[P, Q] \rightsquigarrow \downarrow$ the information gained by the observer $\rightsquigarrow \downarrow$ privacy risk
- $\uparrow D[P, Q] \rightsquigarrow \uparrow$ the information gained by the observer $\rightsquigarrow \uparrow$ benefit of published data

## The Earth Mover Distance (EMD)

- The minimal amount of work needed to transform one distribution to another by moving the distribution mass
- $D[P, Q] = WORK(P, Q, F) = \sum_{i=1}^{m} \sum_{j=1}^{m} d_{ij} f_{ij}$
- $P = (p_1, p_2, ..., p_m)$, $Q = (q_1, q_2, ..., q_m)$
- $d_{ij}$ the ground distance between element $i$ of $P$ and element $j$ of $Q$
- $F = [f_{ij}]$ is the flow of mass from element $i$ of $P$ to element $j$ of $Q$ that minimizes the overall work

| | ZIP Code | Age | Salary | Disease |
|---|---|---|---|---|
| 1 | 476** | 2* | 3K | gastric ulcer |
| 2 | 476** | 2* | 4K | gastritis |
| 3 | 476** | 2* | 5K | stomach cancer |
| 4 | 4790* | $\geq 40$ | 6K | gastritis |
| 5 | 4790* | $\geq 40$ | 11K | flu |
| 6 | 4790* | $\geq 40$ | 8K | bronchitis |
| 7 | 476** | 3* | 7K | bronchitis |
| 8 | 476** | 3* | 9K | pneumonia |
| 9 | 476** | 3* | 10K | stomach cancer |

**Table 4. A 3-diverse version of Table 3**

## $D[P_1, Q]$

- $Q = \{3k, 4k, 5k, 6k, 7k, 8k, 9k, 10k, 11k\}$
- $P_1 = \{3k, 4k, 5k\}$
- If we have an ordered list $\{v_1, v_2 ... v_m\}$ then the ground distance is $\frac{|i-j|}{m-1} = \frac{|i-j|}{8}$
- If we flow $\frac{1}{9}$ mass from $P_1$ to $Q$ as follow $(5k \rightarrow 11k), (5k \rightarrow 10k), (5k \rightarrow 9k), (4k \rightarrow 8k), (4k \rightarrow 7k), (4k \rightarrow 6k), (3k \rightarrow 5k), (3k \rightarrow 4k)$ then $EMD = 1/9 \times (6 + 5 + 4 + 4 + 3 + 2 + 2 + 1)/8 = 0.375$

- A number of privacy notions for protecting data publishing have been proposed
- Nevertheless, data anonymization is still an active research direction
- The trade-off between utility and privacy should be taken into account
- The search is ongoing for measures that scale and maintain probabilistic nature

# Thank you!