Connecting Privacy Models

Synergies between k-Anonymity, t-Closeness and Differential Privacy

Josep Domingo-Ferrer and Jordi Soria-Comas

Universitat Rovira i Virgili, Tarragona

UNESCO Chair in Data Priv@cy

josep.domingo@urv.cat

Outline

1. Improving utility in k-anonymity
2. Differential privacy via t-closeness

Improving utility in k-anonymity

k-Anonymity

Each record must be instinguishable within a group of k records w.r.t. the QI.

- Usually attained through generalization of the QI.
- The variability of the confidential attributes within k-anonymous groups is undefined.
  - too low a variability $\rightarrow$ attribute disclosure
  - too high a variability $\rightarrow$ utility is damaged
- Dealing with attribute disclosure: l-diversity, t-closeness
- No attempts to upper-bound the variability have been made.

Improving utility in k-anonymity

- If the type of data analysis is known before the data release:
  - The k-anonymous groups can be chosen so as to maximize data utility.
- Usually, the kind of analysis is unknown $\rightarrow$
  - No customization of the k-anonymous groups is possible;
  - The lower the variability of the confidential attributes, the better the utility.

Partitioning based on confidential attributes

- To improve utility $\rightarrow$ we minimize the variability of the confidential attributes by
  - Partitioning based on microaggregation of the confidential attributes
- BUT the chance of attribute disclosure is increased, so
  - Use l-diversity or t-closeness to enforce a certain lower bound on the variability in the confidential attributes.

Reducing information loss in QI attributes

- When generating k-anonymous groups by microaggregation of the confidential attributes
  - The homogeneity of the QI within the k-anonymous groups cannot be controlled.
- Replacing the QI in a k-anonymous group with a single generalized value may produce great utility losses.
- To break the relationship between QI and confidential attributes without modifying QIs, split a data set into two tables:
  1. (QI attributes, k-anonymous_group_id)
  2. (k-anonymous_group_id, Confidential attributes)

Differential privacy via t-closeness

t-Closeness

The distribution of the confidential attributes within each k-anonymous group must be similar to the overall distribution of the confidential attributes.

Differential privacy

The presence or absence of any individual should only modify the probability distribution of the output by a small factor.

Microdata releases

- t-Closeness and differential privacy are essentially different.
- For microdata releases, a link between them can be found under some additional assumptions:
  - Only the relation between QI and confidential attributes should be protected.
  - The intruder's prior knowledge about the confidential attributes is limited to their distribution.

Distance function for t-closeness

- t-Closeness depends on a distance function to measure the distance between the global distribution and the distribution associated to each k-anonymous group.
- We use a distance function that closely matches the condition required by differential privacy.

Definition

Let $\mathscr{D}_1$ and $\mathscr{D}_2$ be two random distributions

$$d(\mathscr{D}_1, \mathscr{D}_2) = \max_S \{ \frac{\Pr_{\mathscr{D}_1}(S)}{\Pr_{\mathscr{D}_2}(S)}, \frac{\Pr_{\mathscr{D}_2}(S)}{\Pr_{\mathscr{D}_1}(S)} \}$$

- The distribution of the confidential attributes in a k-anonymous group must differ from the global distribution by a factor not greater than t.

Differential privacy via t-closeness (I)

- With the above assumptions and the proposed distance

$$\exp(\varepsilon)\text{-closeness} \Rightarrow \varepsilon\text{-differential privacy}$$

- Let I be an individual in the data set
- For ε-differential privacy to hold, the inclusion of I in the data set data must affect our knowledge about I's confidential data by a factor not greater than $\exp(\varepsilon)$.

Differential privacy via t-closeness (II)

Intruder's knowledge before accessing the data

The confidential attributes of I are distributed according to the global distribution of the confidential attribute

- From an ε-differential privacy point of view:

Intruder's knowledge from accessing a data set that does not contain I's data
= Intruder's knowledge before accessing the data
= Global distribution of the confidential attribute

Differential privacy via t-Closeness (II)

Intruder's knowledge after accessing the data

The confidential attributes of I are distributed according to the distribution of the corresponding k-anonymous group

- From an ε-differential privacy point of view:

  Intruder's knowledge after accessing the data = Response obtained from a data set that contains I's data = Distribution of the confidential attribute in the corresponding k-anonymous group

- As $\exp(\varepsilon)$-closeness is satisfied with the specified distance, the knowledge gain obtained from including I's data is limited by the multiplicative factor $\exp(\varepsilon)$; hence, ε-differential privacy is satisfied.

Conclusions

- In a general-purpose data release, the variability of the confidential attributes is a key point for data utility.
- Generalization of QI is not optimal in terms of the variability of the confidential attributes.
- We propose a method to attain k-anonymity while minimizing the variability of the confidential attribute.
- Minimizing the variability of the confidential attributes may lead to attribute disclosure.
  - Combine it with l-diversity or t-closeness to guarantee a certain lower bound of variability.
- Under some reasonable conditions t-closeness offers a protection equivalent to that of differential privacy.