Prior k-anonymity via insensitive microaggregation to reduce data utility loss when achieving \( \varepsilon \)-differential privacy in data releases

In insensitive microaggregation

Information loss (k=1 is standard diff. privacy, \( \varepsilon \) from 0.01 to 10)

Construction to achieve \( t \)-closeness and \( \varepsilon \)-differential privacy

Partition of the data set into groups P1, P2, P3... by the quasi-identifiers and bucketization of the confidential attribute to achieve \( t \)-closeness

\[ d(D_1, D_2) = \max_S \left\{ \frac{Pr_{D_1}(S)}{Pr_{D_2}(S)} \times \frac{Pr_{D_2}(S)}{Pr_{D_1}(S)} \right\} \]

where \( D_1 \) and \( D_2 \) are two random distributions differing in one record and \( S \) is an arbitrary set.

◆ The granularity of confidential attribute is reduced, so \( t \)-closeness is achieved with distance

\[ Pr_{D_1}(S) \leq \exp(\varepsilon) \times Pr_{D_2}(S) \]

where \( D_1 \) is the distribution of the confidential attribute in the whole protected data set and \( D_2 \) is the distribution of the confidential attribute in the group \( Pi \) containing a specific individual.

From \( \varepsilon \)-differential privacy to expected \( t \)-closeness

Let \( X \) be an original data set and \( X' \) be a corresponding anonymized data set such that its quasi-identifiers are \( k \)-anonymous and the projection of \( X' \) on the confidential attributes is \( \varepsilon \)-differentially private. Then \( X' \) satisfies expected \( t \)-closeness with

\[ t = g^{-1}(\exp((N - k) \times \varepsilon)) \]

Hence, a greedy way to achieve actual \( t \)-closeness is to keep generating \( \varepsilon \)-differentially private versions of the confidential attribute until a \( t \)-close version is found.

Conclusions

The \( k \)-anonymity, \( t \)-closeness and differential privacy models are connected. Using a prior \( k \)-anonymization step based on insensitive microaggregation allows achieving differential privacy in data set releases with less utility loss. Also, \( \exp(\varepsilon) \)-closeness implies \( \varepsilon \)-differential privacy for uninformed intruders in data releases. Finally, \( k \)-Anonymity for quasi-identifiers combined with \( \varepsilon \)-differential privacy for confidential attributes yields \( t \)-closeness in expectation, with \( t = f(k, \varepsilon) \).

References

