Abstract

Closeness is described as a privacy measure and its advantages are illustrated through examples and experiments on a real dataset. In this paper the closeness can be verified by giving different values for N and T. Government agencies and other organizations often need to publish micro data, e.g., medical data or census data, for research and other purposes. Typically, such data are stored in a table, and each record (row) corresponds to one individual. Generally if we want to publish micro data a common anonymization approach is generalization, which replaces quasi-identifier values with values that are less-specific but semantically consistent. As a result, more records will have the same set of quasi-identifier values. An equivalence class of an anonymized table is defined to be a set of records that have the same values for the quasi-identifiers. To effectively limit disclosure, the disclosure risk of an anonymized table is to be measured. To this end, k-anonymity is introduced as the property that each record is indistinguishable with at least k-1 other records with respect to the quasi-identifier. i.e., k-anonymity requires that each equivalence class contains at least k records. While k-anonymity protects against identity disclosure, it is insufficient to prevent attribute disclosure. To address the above limitation of k-anonymity, a new notion of privacy,
called l-diversity is introduced, which requires that the distribution of a sensitive attribute in each equivalence class has at least l "well represented" values. One problem with l-diversity is that it is limited in its assumption of adversarial knowledge. This assumption generalizes the specific background and homogeneity attacks used to motivate l-diversity. The k-anonymity privacy requirement for publishing micro data requires that each equivalence class contains at least k records. But k-anonymity cannot prevent attribute disclosure. The notion of l-diversity has been proposed to address this; l-diversity requires that each equivalence class has at least l well-represented values for each sensitive attribute. L-diversity has a number of limitations. In particular, it is neither necessary nor sufficient to prevent attribute disclosure. Due to these limitations, a new notion of privacy called "closeness" is proposed.

First the base model t-closeness is presented, which requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table. Then a more flexible privacy model called (n, t)-closeness is proposed. The rationale for using

**References**

- T. Li and N. Li, "Towards Optimal k-Anonymization," Data and

Index Terms

Computer Science
Security

Keywords
Privacy Measure  K-Anonymity  L-Diversity  data Anonymization  (n-t) closeness