K-anonymity

Objective Technique

***k*-anonymity** is a property possessed by certain anonymized data. The concept of *k*-anonymity was first introduced by Latanya Sweeney in a paper published in 2002.

"A release provides k-anonymity protection if the information for each person contained in the release can not be distinguished from at least k-1 individuals whose information also appears on that same release."

The goal is to provide a formal framework for constructing and evaluating algorithms and systems that release information such that the released information limits what can be revealed about properties of the entries that are to be protected.

Problem Definition

"Given person-specific field-structured data, produce a release of the data with scientific guarantees that the individuals who are the subjects of the data cannot be re-identified while the data remain practically useful."

General Methods of Database Security:

1. Statistical Databases: Adding Noise to the data while still maintaining statistical invariant.
2. Multi-Level Databases: Restricting the release of lower level information from which higher level information can not be retrieved. In multi-level databases, each level has different level of security attached to it.
3. Restricting (prohibiting) queries that can reveal sensitive information.
   For example : Table (physician,patient,medication).
   Relations R1(physician,patient) and R2(physician,medication) may not be sensitive.
   However, Relation R(patient,medication) is critically sensitive because medications
   Typically correlate with diseases. Hence, query for this relation can be restricted.

Quasi-identifiers are pieces of information that are not of themselves unique identifiers, but are sufficiently well correlated with an entity that they can be combined with other quasi-identifiers to create a unique identifier.

Definition : k-anonymity

Let $RT(A_1,A_2,.....,A_n)$ be a table and QI_{RT} be the quasi-identifier associated with it. RT is said to satisfy k-anonymity if and only if each sequence of values in $RT[QI_{RT}]$ appears with **at least** k-occurrences in $RT[QI_{RT}]$.

Solution (Advantages)

K-anonymity uses the following two techniques:

1. **Suppression** : In this method, certain values of the attributes are replaced by an asterisk '*'. All or some values of a column may be replaced by '*'. This means that all the information that allows the inference of sensitive information is simply not released.

2. **Generalization** : In this method, individual values of attributes are replaced by with a broader category. For example, the value '19' of the attribute 'Age' may be replaced by ' ≤ 20', the value '23' by '20 < Age ≤ 30' etc.

Example

Name	Age	Gender	State of domicile	Religion	Disease
*	20 < Age ≤ 30	Female	Tamil Nadu	*	Cancer
*	20 < Age ≤ 30	Female	Kerala	*	Viral infection
*	20 < Age ≤ 30	Female	Tamil Nadu	*	TB
*	20 < Age ≤ 30	Male	Karnataka	*	No illness
*	20 < Age ≤ 30	Female	Kerala	*	Heart-related
*	20 < Age ≤ 30	Male	Karnataka	*	TB
*	Age ≤ 20	Male	Kerala	*	Cancer
*	20 < Age ≤ 30	Male	Karnataka	*	Heart-related
*	Age ≤ 20	Male	Kerala	*	Heart-related
*	Age ≤ 20	Male	Kerala	*	Viral infection

This data has **2-anonymity** with respect to the attributes 'Age', 'Gender' and 'State of domicile' since for any combination of these attributes found in any row of the table there are always at least 2 rows with those exact attributes. The attributes available to an adversary are called "quasi-identifiers". Each "quasi-identifier" tuple occurs in at least k records for a dataset with k-anonymity.

Hence, the Lemma
Let $RT(A_1,A_2,.....,A_n)$ be a table, $QI_{RT}=(A_i,...,A_j)$ be the quasi-identifier associated with RT and RT satisfy k-anonymity. Then each sequence of values in $RT[A_x]$ appears with at least k occurrences in $RT[QI_{RT}]$ for x = i,....,j.

Hence, the information available would not match fewer than k individuals. Hence, directly can not be matched to reveal the identity of a particular individual.

Disadvantages (Flaws)

1. Suppression can drastically reduce the quality of data and can sometimes render the data useless.
2. Data precision may be lost during generalization.
3. Prone to attacks which include
    a. Unsorted matching attack against k-anonymity : direct matching will reveal certain information when all the attributes are quasi-identifiers.
    b. Complementary release attack : release of one version after the other
    c. Temporal attacks : release of one version at time T_o and then another version of same table at Time T_t with some changes can also reveal certain sesitive information.

Conclusion

K-anonymity protection model along with its flaws and related attacks have been discussed.