Privacy in Microdata Release

Prof. Ravi Sandhu
Executive Director and Endowed Chair

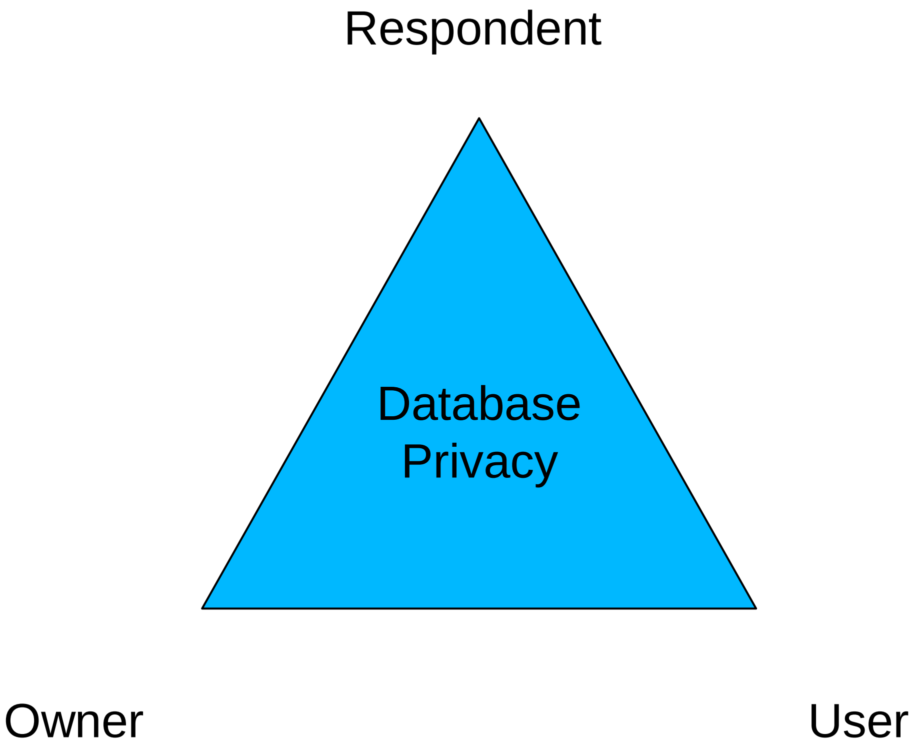
March 22, 2013

ravi.sandhu@utsa.edu
www.profsandhu.com

Topics

- whose privacy?
- k-anonymity
- p-sensitive k-anonymity
- l-diversity
- t-closeness

Domingo-Ferrer 2007



Respondent

Database Privacy

Owner

User

3 independent dimensions

Domingo-Ferrer 2007

PDF 8.1: Page 196 Table 1

Domingo-Ferrer 2008

Definition 1 (k-Anonymity): A protected data set is said to satisfy *k*-anonymity for *k* > 1 if, for each combination of key attributes, at least *k* records exist in the data set sharing that combination.

No attribute confidentiality

Definition 2 (p-Sensitive k-anonymity): A data set is said to satisfy *p*-sensitive *k*-anonymity for *k* > 1 and *p* ≤ *k* if it satisfies *k*-anonymity and, for each group of records with the same combination of key attribute values the number of distinct values for each confidential attribute is at least *p* (within the same group).

Loss of data utility

Example 2: Consider a dataset containing data for 1000 patients. The key attributes in the data set are Age, Height *and* Weight. *There is a single confidential attribute* AIDS *whose values can be ”Yes” or ”No”. Assume that there are only five patients in the dataset with* AIDS=”Yes”. *Imagine that* 2*-sensitive k- anonymity is desired. Clearly, at least one patient with AIDS is needed in each group sharing a combination of key attributes, so that at most five groups can be formed. Therefore, key attributes must be heavily coarsened so that only five combinations of their values subsist.* D

Domingo-Ferrer 2008

Definition 3 (l-Diversity): A data set is said to satisfy *l*-diversity if, for each group of records sharing a combination of key attributes, there are at least *l* "well-represented" values for each confidential attribute.

1. Distinct l-diversity: same as l-sensitivity
2. Entropy l-diversity
3. Recursive (c,l)-diversity

***Skewness attack*.** If, in Example 2, a group has the same number of patients with and without AIDS; in that case, it satisfies distinct 2-diversity, entropy 2-diversity and any recur- sive (*c,* 2)-diversity requirement. However, if an intruder can link a specific patient to that group, that patient can be considered to have 50% probability of having AIDS, in front of 5/1000 for the overall data set.

***Similarity attack*.** If values of a sensitive attribute within a group are l-diverse but semantically similar, attribute disclosure also takes place. E.g. if patients in a 3-diverse data set where Disease is a confidential attribute all have values in {"lung cancer", "liver cancer", "stomach cancer"} an intruder linking a specific individual to that group can infer that the individual has cancer. If the confidential attribute is numerical and values within a group are l-diverse but very similar, the intruder can estimate the confidential attribute value for an individual in that group to a narrow interval.

Domingo-Ferrer 2008

Definition 4 (t-Closeness): *A data set is said to satisfy t-closeness if, for each group of records sharing a combination of key attributes, the distance between the distribution of the confidential attribute in the group and the distribution of the attribute in the whole data set is no more than a threshold t.*

Loss of data utility
enforcing t-closeness destroys the correlations between key attributes and confidential attributes: by definition of t-closeness the values of a confidential attribute have the same distribution for any combination of values of key attributes! The only way to decrease the damage is to increase the threshold t, that is, to relax t-closeness.