Generalizing Data to Provide Anonymity when Disclosing Information (Abstract)

(Paper not available on time)

Pierangela Samarati
SRI International
samarati@csl.sri.com

Latanya Sweeney
Massachusetts Institute of Technology
sweeney@lcs.mit.edu

Abstract

The proliferation of information on the Internet and access to fast computers with large storage capacities has increased the volume of information collected and disseminated about individuals. The existence of these other data sources makes it much easier to re-identify individuals whose private information is released in data believed to be anonymous. At the same time, increasing demands are made on organizations to release individualized data rather than aggregate statistical information. Even when explicit identifiers, such as name and phone number, are removed or encrypted when releasing individualized data, other characteristic data may still uniquely combine into a quasi-identifier, which can allow data recipients to re-identify individuals to whom the data refer.

In this paper, we provide a technique for releasing information such that the ability to link the released data to other information is limited in an attempt to protect the identities of individuals. We use generalization, by which stored values can be replaced with semantically consistent but less precise alternatives, and k-anonymity. A table provides k-anonymity when each combination of values that is associated with a quasi-identifier occurs k or more times, making it difficult to link the quasi-identifier to other data with confidence. We introduce the notions of generalized table and of minimal generalization of a table with respect to a k-anonymity requirement. As an optimization problem, the objective is to minimally distort the data while providing adequate protection. We describe an algorithm that, given a table, efficiently computes a preferred minimal generalization to provide anonymity.