

Injecting Utility into Anonymized Datasets

Daniel Kifer,
Cornell University,
<http://www.cs.cornell.edu/People/dkifer/>

Friday, March 16, 2007
1:30-2:45 PM
TECH Center 111

Abstract:

Limiting disclosure in data publishing is an important issue. As data collection and data management capabilities improve, more entities are collecting and storing detailed information about individuals. Examples include the Census Bureau, insurance companies, credit agencies, retailers, and assorted internet giants. At the same time, these companies face internal and external pressures to provide some data for data mining (e.g., to identify disease outbreaks from medical records, or to study population trends in census data) and research (e.g., to develop better models of individuals' behavior). Thus there are two competing requirements for the released data: it should be useful, and it should preserve the privacy of individuals contained in the data.

In this talk I will overview a novel privacy definition and a common technique for sanitizing the data that gives the necessary privacy guarantees. I will then explain the shortcomings of this approach in terms of the information content of the released data, and show how to increase the utility of the data. Since increased utility is usually accompanied by decreased privacy, I will then show how to reason about the resulting privacy and demonstrate that significantly more information can be revealed while preserving strong privacy guarantees.