

**Department of Applied Mathematics and Statistics
The Johns Hopkins University**

SEMINAR

Robert Garfinkel
Dept. of Operations & Info. Mgmt.
University of Connecticut

October 5, 2006
304 Whitehead Hall
Refreshments: 3:30 p.m.
Seminar: 4:00 p.m.

**STOCHASTIC PROTECTION OF CONFIDENTIAL INFORMATION
IN DATABASES: A HYBRID OF DATA PERTURBATION
AND QUERY RESTRICTION**

ABSTRACT

Data perturbation and query restriction are two of the main methods that have been developed to protect confidential data in statistical database systems. In the former the confidential data is systematically changed to maintain its distributional properties, and to yield answers to queries that are statistically similar to those that would have resulted from the original data. The latter provides exact answers to certain queries but then declines to answer others if the risk of exact disclosure of confidential data becomes too great. We present a new model and corresponding methodology to combine these techniques in such a way that the advantages of both are captured—that is, to provide exact answers to some of the most important queries while answering all others based on the perturbed data and also maintaining the statistical validity of the perturbed data. The model is appropriate, and is shown to be computationally viable for large databases whether the queries are linear or nonlinear. It is shown that an important requirement for preservation of confidentiality is that the two stages provide answers that are consistent with each other. The query restriction phase of the model consists of finding an optimal (based on values of individual queries) subset of a set of important queries to answer exactly without compromising the database. By itself this is an NP-hard problem with a matroid intersection structure that lends itself to an efficient greedy heuristic with well-known performance bounds. Then, given the set of queries that are answered exactly, we show how to implement a data perturbation phase that provides stochastic protection of confidential data and is consistent with these exact answers. Finally, we present computational results on a large database with both linear and nonlinear queries. The results indicate that very many queries can be answered exactly and further the proposed perturbation approach provides more accurate answers than the standard perturbation approach. Together these results highlight the practical viability of the proposed techniques.

(This is joint work with Manuel Nunez and Ram Gopal.)